The Structure of Mathematical Expressions

An arXiv Case Study

Deyan Ginev and Bruce R. Miller

National Institute of Standards and Techonology

May 4, 2012

Contents

Chapter 1

Introduction

In this study, we survey the notational diversity of present-day mathematical expressions, in order to uncover their linguistic phenomena. A practical motivation for this study is to provide a foundation for determining the boundary between syntactic and semantic phenomena in said expressions, from the perspective of language modeling. The ultimate goal of this project is to construct a grammar of mathematical expressions, which captures all relevant syntactic properties established in this study, and allows for the semantic analysis necessary to model and observe the semantic relationships.

1.1 Motivation

We want to enable machine-reading of formulas, in order to provide a variety of user-assistance services, such as semantic search, text-to-speech synthesis, semantic interactions (definition lookup), as well as computer algebra support ("evaluate subexpressions on demand") and ultimately computer verification ("does that proof step really hold?").[1] EdN:1

1.2 Related Resources

Notation census, beginnings of study are in Deyan's thesis, Naproche and FMathL have examples, but no real systematic study.[2] EdN:2

[1] EdNote: expand

[2] EdNote: expand

Chapter 2

Methods

2.1 Training Corpus

The primary corpus on which we base this investigation is the Cornel pre-print archive “arXiv”[3], consisting of over 700,000 articles in 37 scientific subfields. EdN:3

arXiv Sandbox

[4] EdN:4

As a secondary resource, we we will also consult entry-level literature on highschool mathematics, in order to exhibit basic phenomena, as well as to demonstrate phenomena apriori known to the authors.[5] EdN:5

2.2 Structural Annotation

As one of the goals of our study is to establish a first guess of an underspecified operator tree[6], EdN:6 any annotation must at its core mark up the applicative logical structure of the mathematical expression. This process will build up a formula tree, the collection of which can later be used as a gold standard for developing a grammatical model of the language of symbolic mathematics.

[7] [8] EdN:7 EdN:8

2.3 Annotation Vocabulary

Another core goal is to discover and describe interesting linguistic phenomena that occur naturally in our corpus. Examples of what we consider “interesting” are phenomena that

[3] EdNote: cite here

[4] EdNote: Say that, on the arXiv front, we first start with the train sandbox from Deyan's thesis

[5] EdNote: Wikipedia? PEMDAS?

[6] EdNote: make sure the concepts are introduced and/or rephrase

[7] EdNote: I'm currently thinking of rendering the annotations as trees (tikz,pstricks...custom tree drawing package?), so that the annotator can proofread the annotations in an intuitive manner.

[8] EdNote: In the XHTML, I'm thinking of ContentMML+SVG rendering, all of this figured out by the binding, maybe a custom stylesheet?

Train1	Differential Geometry `http://arxmliv.kwarc.info/files/9609/dg-ga.9609012`
Train2	Quantum Physics `http://arxmliv.kwarc.info/files/0910/0910.5733/`
Train3	High Energy Physics - Theory `http://arxmliv.kwarc.info/files/9407/hep-th.9407125/`
Train4	Commutative Algebra `http://arxmliv.kwarc.info/files/0809/0809.4873/`
Train5	Statistics Theory `http://arxmliv.kwarc.info/files/0905/0905.1486/`
Train6	General Relativity and Quantum Cosmology `http://arxmliv.kwarc.info/files/0807/0807.2507/`
Train7	Cosmology and Extragalactic Astrophysics `http://arxmliv.kwarc.info/files/0908/0908.2548`
Train8	Exactly Solvable and Integrable Systems `http://arxmliv.kwarc.info/files/0905/0905.2033`
Train9	Geometric Topology `http://arxmliv.kwarc.info/files/0809/0809.4477`
Train10	Algebraic Geometry `http://arxmliv.kwarc.info/files/0704/0704.0537`

Table 2.1: Sandbox of Ten Random arXiv Papers from Diverse Scientific Subfields

induce ambiguity, or legitimize what would typically be ungrammatical fragments. Cases of ambiguity are well-known to follow from semantic overloading of symbols, implicit argument scopes of operations or eliding syntax, leaving the reader with the task of guessing the "invisible" dynamics.Use of custom shorthands, however, as well as custom notations in general, expands the grammar of symbolic mathematics, often in completely non-standard ways that can only be grasped through a deep understanding of the document at hand.

As multiple interesting observations can be made for a single large mathematical formula, it is natural to annotate multiple relevant subexpressions. More concretely, for each phenomenon of interest, we annotate the greatest common subtree (GCT) of all participating subtrees. In case we find a long-range relationship in a large formula, the annotation would hence be placed on the formula root.

The annotations can be utilized for different purposes - browsing by specific phenomena, syntactic feature or lemma, training a classifier, etc. Thus, we take a compositional, standardized approach to providing labels from a fixed vocabulary for the relevant ontological classes of structural properties.

EdN:9 [9]

[9] EdNote: Additional tokens: super, sub, fenced

Property	**Keywords**
Fixity	over, under, prefix, infix, postfix, superfix, subfix, circumfix, transfix, nofix[1]
Role (Symbols)	separator, modifier, relation, operator, metarelation, binder
Role (Objects)	factor, term, statement, variable, constant, modified
Role (Structure)	tuple, sequence, expression, shorthand, template, language
Composition	invisible, atom, complex, chained
Shallow Semantics	type, function, constructor, other
Linguistic	ellipsis, metonymy, ambiguity, vagueness, anaphora
Math Practices	framing

Table 2.2: Keyword Vocabulary for Syntactic Properties

Chapter 3

A Study of Mathematical Syntax

3.1 Basics

Foundations

[10] [11] [12]

EdN:10
EdN:11
EdN:12

High School

[13] [14]

EdN:13
EdN:14

3.2 Discrete math

Set Theoretic Notations

[15] [16]

EdN:15
EdN:16

Logical Operators

[17]

EdN:17

Combinatorics

[18] [19]

EdN:18
EdN:19

[10] EdNote: arithmetic, grouping fences and equality

[11] EdNote: basic relations and orderings

[12] EdNote: arithmetic and algebraic sequences?

[13] EdNote: geometry here, otherwise a separate geometry subsection

[14] EdNote: trigonometry, complex and rational numbers

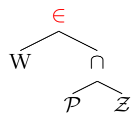
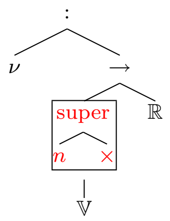
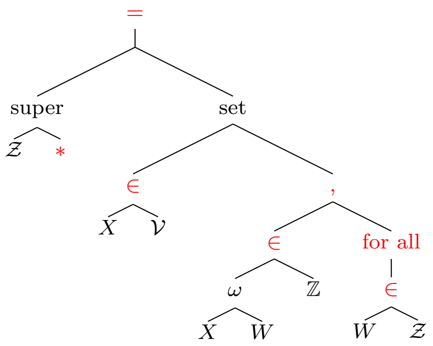
[15] EdNote: elementhood, inclusions, set constructors, overloaded arith ops

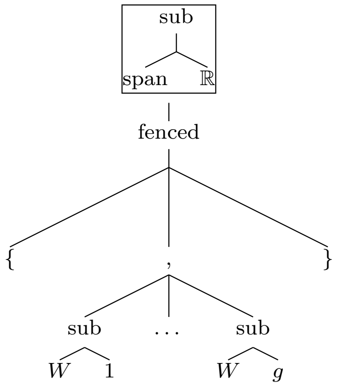
[16] EdNote: also maps : domains -¿ codomains, xRy notations

[17] EdNote: classic logic, HOL, type theories

[18] EdNote: Infinite sums

[19] EdNote: binomials, combinations, permutations,

	Expression	Denotation	Annotation
1.	$W \in \mathcal{P} \cap \mathcal{Z}$	set membership	
	Discussion: set operators precede set relations, [Train1]		
2.	$\nu : \overset{n}{\times}\mathbb{V} \longrightarrow \mathbb{R}$	a map	 : ν → super n × $\mathbb{R}$ $\mathbb{V}$
	Discussion: n-ary cross-product, [Train1]		
3.	$\mathcal{Z}^* = \{X \in \mathcal{V} \mid \omega(X, W) \in \mathbb{Z},\ \text{for all } W \in \mathcal{Z}\}$	definition to set	 = super set $\mathcal{Z}$ * ∈ , X $\mathcal{V}$ ∈ for all ω $\mathbb{Z}$ ∈ X W W $\mathcal{Z}$
	Discussion: NL mixins, quantified relation, [Train1]		

4.	$\text{span}_{\mathbb{R}}\{W_1, \ldots, W_g\}$	span of a set	 sub
			span
			$\mathbb{R}$
			fenced
			{
			,
			}
			sub
			...
			sub
			W
			1
			W
			g

	Discussion: set operators can take fenced yet not simply *grouped* arguments, [Train1]		

Table 3.1: Set Theory Notations, Part 1

Number Theory

[20] [21] [22] [23]

EdN:20
EdN:21
EdN:22
EdN:23

Graph Theory

[24] [25] [26]

EdN:24
EdN:25
EdN:26

Algebra

[27] [28] [29] [30]

EdN:27
EdN:28
EdN:29
EdN:30

Functions Theory

[31]

EdN:31

3.3 Continuous math

Calculus

[32]

EdN:32

Probability

[33] [34]

EdN:33
EdN:34

Interval Notation and Arithmetic

[35]

EdN:35

Topology

[36]

EdN:36

[20] EdNote: modulo modifiers

[21] EdNote: tuples

[22] EdNote: divisibility notations $a \mid b$ and b/a

[23] EdNote: DLMF sneaky notations

[24] EdNote: edge and vertex notations

[25] EdNote: incidence and adjacency notations

[26] EdNote: Wiki is very nice: `http://en.wikipedia.org/wiki/Glossary\_of\_graph\_theory`

[27] EdNote: vectors

[28] EdNote: maps and complements

[29] EdNote: groups

[30] EdNote: lattices

[31] EdNote: talk about associativity of application and composition, ";" and "$\circ$" as notation variants, discuss complex examples

[32] EdNote: differentials, integrals, limits, remember brownian motion integral notations!

[33] EdNote: Bayes formula with multiple denotations of P

[34] EdNote: Various conditional and joint probability notations

[35] EdNote: introduce interval notations, then move to interval arithmetic

[36] EdNote: manifold constructors and notations

Differential Geometry

Some intro text?

EdN:37 [37]

EdN:38 [38]

[37] EdNote: more on $\pmod{x}$ notations

[38] EdNote: Complex named enttity: "$U(1)$ Chern-Simons gauge theory."

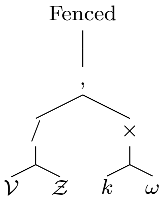
	Expression	Denotation	Annotation
1.	$(\mathcal{V}/\mathcal{Z}, k\omega)$	symplectic torus	 Fenced , / × $\mathcal{V}$ $\mathcal{Z}$ k ω
	Discussion: [Train1]		
2.	$\mathcal{Z}$	self-dual lattice	atom
	Discussion: [Train1]		
3.	$(\mathcal{V}, \omega)$	symplectic vector space	circumfix constructor
	Discussion: [Train1]		
4.	$Lag(\mathcal{V})$	Lagrangian Grassmannian	circumfix constructor
	Discussion: [Train1]		
5.	$Lag_4(\mathcal{V})$	4-fold covering space	prefix constructor
	Discussion: [Train1]		
6.	$\mathcal{M}_\Sigma$	moduli space	subfix constructor
	Discussion: [Train1]		
7.	Σ	Riemann surface	atom variable
	Discussion: [Train1]		
8.	$H^1(\Sigma; \mathbb{R})$	chomology space	transfix constructor
	Discussion: [Train1]		
9.	$H^1(\Sigma; \mathbb{R})/H^1(\Sigma; \mathbb{Z})$	torus	infix operator
	Discussion: [Train1]		
10.	(M, ω)	symplectic manifold	circumfix constructor
	Discussion: [Train1]		
11.	$f \in \mathcal{C}^\infty(M)$	smooth function	atom modified
	Discussion: [Train1]		
12.	X_f	field	subfix constructor
	Discussion: [Train1]		
13.	$\lrcorner$	interior product	complex infix operator
	Discussion: Formed via \mathop in TeX, [Train1]		
14.	$[\omega] \in H^2(M; \mathbb{R})$	cohomology class	complex variable modified
	Discussion: [Train1]		
15.	$(\cdot, \cdot)$	template, hermitian metric	template tuple
	Discussion: [Train1]		

Table 3.2: Differential Geometry Notations, Part 1

	Expression	Denotation	Annotation
16.	$-2\pi\mathrm{i}\,\omega$	complex number	invisible prefix infix operator expression
	Discussion: [Train1]		
17.	$(\mathcal{L}, \nabla)$	prequantum line bundle	circumfix constructor
	Discussion: [Train1]		
18.	$U \subset M$	open subset	modified atom
	Discussion: [Train1]		
19.	$\mathcal{L}\big\|_U$	restricted line bundle	modified atom
	Discussion: postfix restriction via "$\big\|_U$", [Train1]		
20.	$s \in \Gamma(U; \mathcal{L})$	nonzero section	modified atom
	Discussion: [Train1]		
21.	$\nabla s = -2\pi\mathrm{i}\,\theta\, s$	equation	relation
	Discussion: [Train1]		
22.	$\omega\big\|_U = d\theta$	equation	relation
	Discussion: [Train1]		
23.	$T_x M$	bundle	metonymy infix invisible constructor
	Discussion: metonymy for tangent bundle, concat is a space-constructor, [Train1]		
24.	$\omega\big\|_{\mathcal{P}_x} \equiv 0$	equivalence	relation
	Discussion: [Train1]		
25.	$\dim \mathcal{P}_x = \frac{1}{2} \dim T_x M$	equality	relation
	Discussion: dim has lower precedence than invisible bundle-formation, [Train1]		
26.	$[X, Y] \in \mathcal{X}_{\mathcal{P}}(M)$	commutator is in set	relation
	Discussion: used as verb phrase in sentence, [Train1]		
27.	$\nabla^{\mathcal{P}}$	covariant differentiation	scripted prefix op
	Discussion: big op?, [Train1]		
28.	$\begin{array}{rll} \nabla^{\mathcal{P}} : \mathcal{X}_{\mathcal{P}}(M) \times \mathcal{X}_{\mathcal{P}}(M) & \longrightarrow & \mathcal{X}_{\mathcal{P}}(M) \\ (X, Y) & \longmapsto & \nabla^{\mathcal{P}}_X Y\,, \end{array}$	domain specification	type modifier
	Discussion: alignment splits type statement, trailing comma [Train1]		
29.	$(\nabla^{\mathcal{P}}_X Y) \lrcorner\, \omega = X \lrcorner\, d\,(Y \lrcorner\, \omega)\,.$	definitional assignment	infix relation
	Discussion: trailing dot, [Train1]		
30.	$\Pi_{\mathcal{P}} : M \to M/\mathcal{P}$	canonical projection map	type modifier
	Discussion: [Train1]		

Table 3.3: Differential Geometry Notations, Part 2

	Expression	Denotation	Annotation
31.	T^g	g-dimensional torus	complex object
	Discussion: script means dimensionality[Train1]		
32.	$q_1, \dots, q_g$	coordinate functions	enumerative sequence
	Discussion: [Train1]		
33.	$X_{q_1}, \dots, X_{q_g}$	Hamiltonian vector fields	enumerative sequence
	Discussion: [Train1]		
34.	$q_1 \circ \Pi_{\mathcal{P}}, \dots, q_g \circ \Pi_{\mathcal{P}}$	functions	enumerative sequence
	Discussion: sequence elements are applicative objects, [Train1]		
35.	$\gamma_1(\Lambda), \dots, \gamma_g(\Lambda)$	basis for a homology group	enumerative sequence
	Discussion: [Train1]		
36.	$j_i(y) = \int\limits_{\gamma_i(\Lambda)} \theta,\ \text{where } y = \Pi_{\mathcal{P}}(\Lambda)\,,$	definitional assignment	infix relation
	Discussion: integral has no binder, nat. lang. modifier, punctuation, [Train1]		
37.	$\mathrm{Det}\,\mathbb{V} = \overset{n}{\wedge}\mathbb{V}$	definitional assignment	infix relation
	Discussion: n-ary wedge?, hidden binder on $\mathbb{V}$, [Train1]		
38.	$\kappa\,(X_{j_1}\big\vert_{\Lambda}, \dots, X_{j_g}\big\vert_{\Lambda}) = 1\,.$	canonically defined density	infix relation
	Discussion: bars as postfix, within a sequence [Train1]		
39.	$(\nabla^{\mathcal{P}}_{W}\,\nu)(X^*_1, \dots, X^*_g) \;=\; W(\,\nu(X^*_1, \dots, X^*_g)\,)\,,$	definitional assignment[1]	infix relation
	Discussion: applied function is fenced, [Train1]		
40.	$0 \longrightarrow \Omega^0_{\mathcal{P}}(\mathcal{L}_{\mathcal{P}}) \xrightarrow{\nabla^{\mathcal{P}}} \Omega^1_{\mathcal{P}}(\mathcal{L}_{\mathcal{P}}) \xrightarrow{\nabla^{\mathcal{P}}} \cdots \xrightarrow{\nabla^{\mathcal{P}}} \Omega^g_{\mathcal{P}}(\mathcal{L}_{\mathcal{P}}) \longrightarrow 0$	complex	type?
	Discussion: arrows as transitions, ellipsis, [Train1]		
41.	$\overset{k}{\wedge}\mathcal{P}^* \otimes \mathcal{L}_{\mathcal{P}}$	line bundle	applicative constructor?
	Discussion: which operator binds first?,[Train1]		
42.	$c_\Lambda = \int_\Lambda f_\Lambda \hat{\kappa}$	definitional assignment	infix relation
	Discussion: bound variable in integral subscript[Train1]		
43.	$H^g(M; \mathcal{P}, \mathcal{L}_{\mathcal{P}}) \;\cong\; \bigoplus\limits_{\Lambda \subset \mathcal{BS}_{\mathcal{P}}} S_\Lambda$	natural isomorphism	infix relation
	Discussion: n-ary $\oplus$, congruence, [Train1]		
44.	(s, s')	function on Λ	circumfix constructor
	Discussion: shorthand constructor for a function, [Train1]		
45.	$\int\limits_{\Lambda}(s, s')\,\mu * \mu'$	integral application	prefix application
	Discussion: binder in subscript, infix operator "$*$" binds stronger than invisible apply [Train1]		

Table 3.4: Differential Geometry Notations, Part 3

	Expression	Denotation	Annotation
46.	$\langle\langle\cdot,\cdot\rangle\rangle : \mathcal{H}_{\mathcal{P}_2} \times \mathcal{H}_{\mathcal{P}_1} \to \mathbb{C}$	sesquilinear pairing pattern	type modifier
	Discussion: operator pattern, along with operator type [Train1]		
47.	$\omega \;=\; \sum_{i=1}^{g} dp^i \wedge dq^i$	symplectic form	infix relation
	Discussion: sum over wedge applications, [Train1]		
48.	$p^i = \text{constant}$		
	Discussion: bad text/math modality, RHS outside of math[Train1]		
49.	$W \cdot (X, \lambda) \;=\; (X + W, \epsilon(W)\, \mathrm{e}^{\pi i k \omega(W,X)}\, \lambda),$	$\mathscr{Z}$-action definition	infix relation
	Discussion: defines operator $\cdot$, arguments quantified via NL following the math expression,[Train1]		
50.	$l\,(l \leq g)$	dimension	modified object
	Discussion: invisible modifier, using fenced relation,[Train1]		
51.	$(W_1, \ldots, W_g; W_1^{\perp}, \ldots, W_g^{\perp})$	symplectic basis	circumfix constructor
	Discussion: distinction between commas and semicolon, [Train1]		
52.	$i = 1, \ldots, g$	definitional range	infix relation
	Discussion: defined to be a sequence? or modifying restriction over a range?, [Train1]		
53.	$X \in \mathcal{V}$, $W \in \mathcal{P}$	conjunction of statements	sequence of relations
	Discussion: comma denotes NL "and" between two relational statements,[Train1]		
54.	$k\,\omega(W_i, X) \in \mathbb{Z}, \qquad i = 1, \ldots, g\,,$	statement	modified relation
	Discussion: four scopes of commas, also hinted by spacing, [Train1]		
55.	$q_i \pmod{k}$	modulo	invisible modifier
	Discussion: fenced modifier argument, prefix mod?,[Train1]		
56.	$\Lambda_{\mathbf{q}=(q_1,\ldots,q_g)} : k\,\omega(W_i, X) \;=\; q_i \pmod{k},\, i = 1, \ldots, g.$	orbit description	infix relation
	Discussion: complex expression, rich in phenomena,[Train1]		
57.	$\hat{\Lambda}_{\mathbf{q},\mathbf{l}} \;=\; \{X \in \mathcal{V} \mid\, k\,\omega(W_i, X) = q_i + kl_i,\; i = 1, \ldots, g\}.$	definitional assignment	infix relation
	Discussion: modified relation!,[Train1]		
58.	$\{\sigma_{\mathbf{q}} = s_{\mathbf{q}} \otimes \delta_{\mathbf{q}}\}_{\mathbf{q}\in(\mathbb{Z}/k\mathbb{Z})^g}$	standard unitary basis if $\mathcal{H}_{\mathcal{P}}$	set constructor
	Discussion: relational modifier to set constructor argument, subscripted set range, [Train1]		
59.	$k\omega(W_i, T_j) \;=\; \delta_{ij} \qquad i, j = h + 1, \ldots, g$	statement	infix relation
	Discussion: spaces determine equality scopes, act as conjunctions; equality on sequence and range[Train1]		
60.	$[\mathbf{l}] = [(l_1, \ldots, l_g)] \in \mathbb{Z}^g/\omega(2,1)\mathbb{Z}^g$	equivalence class	doubly modified object
	Discussion: two relations modify $[\mathbf{l}]$, chained modifying? or nested modifying?[Train1]		

Table 3.5: Differential Geometry Notations, Part 4

	Expression	Denotation	Annotation
61.	${}^{t}\mathbf{q}_1$	?	scripted atom
	Discussion: prescript t, but what does it mean?, [Train1]		
62.	$\sum_{\substack{\mathbf{q}_2\\0\leq q_{2i}\leq k\lvert\det\omega(2,1)\rvert-1}} \sum_{[\mathbf{l}],[\mathbf{l}']} \cdots$	nested summation	prefix operator apply
	Discussion: stacked subscripts, and subscript sequence[Train1]		
63.	$\oplus_{i=1}^{g}\mathbb{Z}W_{1i} \oplus \oplus_{i=1}^{g}\mathbb{Z}W_{1i}^{\perp}$	?	infix apply
	Discussion: mixing prefix n-ary $\oplus$ with infix binary $\oplus$.[Train1]		
64.	$\tau : \mathit{Lag}(\mathcal{V}) \times \mathit{Lag}(\mathcal{V}) \times \mathit{Lag}(\mathcal{V}) \to \mathbb{Z}$	function declaration	typed modifier
	Discussion: $\times$ is weaker than invisible application, when in a typing context?, [Train1]		
65.	$L_1, L_2, L_3, L_4 \in \mathit{Lag}(\mathcal{V})$	set membership	infix relation
	Discussion: multirelation? membership holds for each of the sequence entries on LHS, [Train1]		
66.	$r \equiv \pmod{q}$	equivalence	infix relation
	Discussion: second-order relation, r is used as a superscript notation for modulo apply, [Train1]		
67.	$U_{\mathcal{P}}(b) = F_{\mathcal{P},b\mathcal{P}} \circ b \,:\, \mathcal{H}_{\mathcal{P}} \to \mathcal{H}_{\mathcal{P}}$	unitary operator definition	typed modifier
	Discussion: modifier for assignment, followed by modifier for type, [Train1]		
68.	$(W_i' = bW_i\,;W_i'^{\perp} = bW_i^{\perp})$	symplectic basis	circumfix constructor
	Discussion: two modified arguments to the main constructor, [Train1]		
69.	$(x,c) \in \mathcal{T} \times Sp(2g,\mathbb{R}) \xrightarrow{b} (b(x),bc) \in \mathcal{T} \times Sp(2g,\mathbb{R})$	left action	arrow transition
	Discussion: $\times$ is not used in a typing sense, but in a cross-product sense, inducing a different arrow interplay [Train1]		
70.			
	Discussion: [Train1]		
71.			
	Discussion: [Train1]		
72.			
	Discussion: [Train1]		
73.			
	Discussion: [Train1]		
74.			
	Discussion: [Train1]		
75.			
	Discussion: [Train1]		

Table 3.6: Differential Geometry Notations, Part 5

EdN:39
EdN:40
[39] [40]

3.4 Other fields

Quantum Physics

EdN:41
EdN:42
[41] [42] ⋮

[39] EdNote: Scripts give you new names or new objects

[40] EdNote: Prime scripts can be used for both naming and operating

[41] EdNote: Bra-ket notation

[42] EdNote: computer science, biology, chemistry...

Chapter 4

Discussion

Chapter 5

Conclusion