

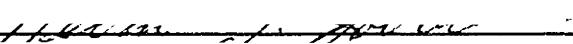
Designing Molecules Possessing Desired Physical Property Values  
Volume 1

by  
Kevin G. Joback  
B.E. Stevens Institute of Technology 1982  
S.M. Massachusetts Institute of Technology 1984

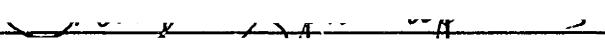
Submitted to the Department of  
Chemical Engineering  
in Partial Fulfillment of the  
Requirements of

Doctor of Philosophy at the Massachusetts Institute of Technology

© Massachusetts Institute of Technology 1989  
© Kevin G. Joback 1989

Signature of Author 

Department of Chemical Engineering  
June 27, 1989

Certified by 

Professor George Stephanopoulos  
Thesis Supervisor

Accepted by 

Vol. 1  
MASSACHUSETTS INSTITUTE  
OF TECHNOLOGY

OCT 3 1989

LIBRARIES

William M. Deen  
Chairperson, Departmental Graduate Committee

# Abstract

The search for new compounds possessing unique and improved properties is an important aspect of chemical engineering. Much of this search is conducted at the experimental level. Chemists synthesize and test thousands of compounds searching for a new pharmaceutical and tens of thousands searching for a new insecticide. Computer implementation of techniques which determine physical and chemical properties now allow much of this search to be performed at a computational rather than experimental level.

Numerous techniques exist for estimating a compound's physical property values given its molecular structure. My thesis research has focused on using these estimation techniques in the inverse manner: specifying desired physical property values and designing the molecular structure of compounds which possess these values. I present a methodology which performs such a molecular design.

The methodology consists of six steps:

1. **Problem Formulation:** problem formulation involves determining constraints on important physical properties.
2. **Target Transformation:** equation oriented physical property estimation techniques are used to propagate these constraints to constraints on properties estimated by group contribution techniques.
3. **Group Selection:** the heart of the methodology consists of two procedures: one interactive and one automatic. Both are based upon group contribution estimation techniques. The interactive procedure uses interactive graphics. Visually representing the problem allows a molecular designer to use his or her own knowledge to guide the search. The automatic procedure uses a hierarchical generate and test paradigm to exhaustively and efficiently search a large number of molecules.
4. **Molecule Enumeration:** the design procedures produce collections of groups. These groups can be connected in several ways to produce complete molecules. All possible ways of connecting these groups are enumerated giving all possible molecular structures.
5. **Molecule Screening:** the design procedures utilize knowledge at the group level of detail. After enumeration we have complete molecules and thus can use molecule level knowledge. This knowledge is in the form of rules specifying molecular substructures which are not allowed in the designed molecules.

6. **Final Evaluation:** it is sometimes necessary to use simplified physical property estimation models to design molecules. The final step of the methodology is to further prune the designed molecules using very accurate estimation techniques.

My thesis presents these six methodological steps. Case studies in refrigerant design, polymer design, solvent design, and drug design demonstrate the methodology.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Objective . . . . .	3
1.2	Incentive . . . . .	3
1.3	Overview . . . . .	4
1.4	Scope . . . . .	6
<b>2</b>	<b>Previous Work</b>	<b>8</b>
2.1	Estimating Physical Properties . . . . .	8
2.1.1	Pattern Recognition . . . . .	9
2.1.2	Topological Techniques . . . . .	15
2.1.3	Group Contribution Techniques . . . . .	16
2.1.4	Equation Oriented Techniques . . . . .	19
2.1.5	Molecular Modeling Techniques . . . . .	20
2.2	Selecting Chemical Products . . . . .	23
2.2.1	Godfrey . . . . .	24
2.2.2	Francis . . . . .	25
2.2.3	Berg . . . . .	25
2.3	Designing Chemical Products . . . . .	27
2.3.1	Solvent Design . . . . .	28
2.3.2	Polymer Design . . . . .	30
2.3.3	Polymer Coatings Design . . . . .	33
2.3.4	Drug Design . . . . .	36
2.4	Generate and Test . . . . .	42

2.5 DENDRAL . . . . .	43
2.6 Interval Arithmetic . . . . .	49
<b>3 Modeling Molecular Design</b>	<b>51</b>
3.1 Constraint Elucidation . . . . .	51
3.2 Property Estimation . . . . .	52
3.3 Molecule Generation . . . . .	52
3.4 Molecule Enumeration . . . . .	53
3.5 Detailed Evaluation . . . . .	53
3.6 My Methodology . . . . .	54
<b>4 Problem Formulation</b>	<b>57</b>
4.1 Solvent Design . . . . .	58
4.2 Refrigerant Design . . . . .	62
4.2.1 Graphical Constraints . . . . .	66
4.3 Barrier Polymer Design . . . . .	68
4.4 Sources of Constraints . . . . .	70
4.4.1 Constraints from Equipment . . . . .	70
4.4.2 Storage . . . . .	71
4.4.3 Physical Property Constraints . . . . .	72
4.4.4 State . . . . .	72
4.5 Summary . . . . .	72
<b>5 Target Transformation</b>	<b>73</b>
5.1 Estimation Procedures . . . . .	73
5.2 Selection Criteria . . . . .	75
5.3 Restriction . . . . .	75
5.4 Group Consistency . . . . .	76
<b>6 Automatic Design</b>	<b>79</b>
6.1 Generate and Test . . . . .	80
6.2 The Generator . . . . .	80

6.2.1	Molecules	82
6.3	The Tester	83
6.3.1	Property Constraints	83
6.3.2	Structural Constraints	84
6.3.3	Chemical Constraints	91
6.4	Algorithm	92
6.5	A Combinatorial Explosion	94
6.6	Meta-Groups	95
6.7	Meta-Molecules	95
6.8	Meta-Contributions	97
6.8.1	Intervals	97
6.9	Meta-Properties	98
6.9.1	Excess Width	100
6.9.2	Causes of Excess Width	101
6.10	Meta-Algorithm	103
6.10.1	Division Strategies	108
6.11	Algorithm Evaluation	113
<b>7</b>	<b>Interactive Design</b>	<b>130</b>
7.1	Procedure Basis	131
7.2	Constraint Visualization	145
7.3	Interactive Pruning	148
7.4	Cognitive Model of Interactive Design	151
7.4.1	Representation	151
7.4.2	Focus of Attention	154
7.4.3	Pattern Recognition	154
7.4.4	Zooming	155
7.4.5	Design Facilities	155
7.4.6	Cognitive Sample	156

<b>8 Enumeration</b>	<b>166</b>
8.1 The Problem . . . . .	167
8.2 Combinatorics . . . . .	168
8.3 Enumeration Procedure . . . . .	170
8.4 Implementation Difficulties . . . . .	176
<b>9 Molecule Screening</b>	<b>178</b>
9.1 The Problem . . . . .	178
9.2 Disallowed Substructures . . . . .	179
9.3 Substructure Representation . . . . .	180
9.4 Substructure Identification Procedure . . . . .	183
9.4.1 Stage 1: Matching Atom and Bond Types . . . . .	184
9.4.2 Stage 2: Matching Atoms and Bonds . . . . .	184
9.4.3 Stage 3: Match Individual Atom–Bond Connections . . . . .	185
9.4.4 Stage 4: Match All Atom–Bond Connections . . . . .	186
9.5 Group Formation . . . . .	189
<b>10 Molecule Evaluation</b>	<b>190</b>
<b>11 Refrigerant Design</b>	<b>192</b>
11.1 Refrigeration . . . . .	192
11.2 Current Refrigerants . . . . .	194
11.3 Problems with Chlorofluorocarbons . . . . .	195
11.4 Problem: Automotive Air Conditioning . . . . .	198
11.5 Freon 12 . . . . .	198
11.6 Problem Formulation . . . . .	198
11.7 Target Transformation . . . . .	200
11.7.1 Transformation for Interactive Design . . . . .	201
11.7.2 Resulting Properties and Groups . . . . .	203
11.7.3 Transformation for Automatic Design . . . . .	203
11.7.4 Resulting Properties and Groups . . . . .	205

11.8 Interactive Design . . . . .	206
11.8.1 Interactive Results . . . . .	207
11.8.2 Replacing Chlorine . . . . .	213
11.9 Automatic Design . . . . .	213
11.9.1 Automatic Results . . . . .	216
<b>12 Polymer Design</b>	<b>221</b>
12.1 IC Packaging . . . . .	222
12.2 Current Encapsulants . . . . .	222
12.3 Problem Formulation . . . . .	223
12.4 Target Transformation . . . . .	224
12.4.1 Transformation for Automatic Design . . . . .	225
12.4.2 Consistent Groups . . . . .	227
12.5 Automatic Results . . . . .	229
<b>13 Solvent Design</b>	<b>235</b>
13.1 The Problem . . . . .	235
13.2 Current Solvents . . . . .	237
13.3 Problem Formulation . . . . .	239
13.4 Target Transformation . . . . .	240
13.4.1 Group Contributions . . . . .	246
13.5 Interactive Design . . . . .	247
13.5.1 Homologous Series . . . . .	250
13.5.2 Solvent Mixtures . . . . .	250
13.6 Automatic Design using UNIFAC . . . . .	254
<b>14 Drug Design</b>	<b>256</b>
14.1 Steps in Drug Design . . . . .	257
14.1.1 Target Biological Properties . . . . .	257
14.1.2 Identification of “Lead” Compounds . . . . .	258
14.1.3 Analog Synthesis and Model Development . . . . .	258

14.1.4 Optimization of Drug Potency . . . . .	261
14.2 The Problem . . . . .	262
14.3 Problem Formulation . . . . .	263
14.4 Target Transformation . . . . .	266
14.5 Interactive Design . . . . .	267
<b>15 Conclusions</b>	<b>270</b>
<b>16 Recommendations</b>	<b>272</b>
16.1 Structural Constraints . . . . .	273
16.2 Meta-Group Strategies . . . . .	273
16.3 Design Methodology Cooperation . . . . .	273
16.4 Specific Estimation Techniques . . . . .	274
16.5 Molecular Display . . . . .	274
<b>A Estimation Techniques</b>	<b>289</b>
A.1 Normal Boiling Point . . . . .	289
A.1.1 Group Contribution Technique . . . . .	291
A.1.2 Equation Oriented Technique . . . . .	291
A.2 Reduced Boiling Point . . . . .	293
A.2.1 Group Contribution Technique . . . . .	294
A.3 Critical Temperature . . . . .	294
A.3.1 Equation Oriented Technique . . . . .	296
A.4 Critical Pressure . . . . .	296
A.4.1 Group Contribution Technique . . . . .	297
A.4.2 Equation Oriented Technique . . . . .	297
A.5 Vapor Pressure . . . . .	299
A.5.1 Equation Oriented Technique . . . . .	300
A.6 Acentric Factor . . . . .	301
A.6.1 Equation Oriented Technique . . . . .	301
A.7 Enthalpy of Vaporization at $T_b$ . . . . .	302

A.7.1	Group Contribution Technique	302
A.7.2	Equation Oriented Technique	305
A.8	Enthalpy of Vaporization	305
A.8.1	Equation Oriented Technique	305
A.9	Factors	306
A.9.1	$F_1$ Group Contribution Technique	306
A.9.2	$F_2$ Group Contribution Technique	308
A.9.3	$F_2$ Assumption	310
A.9.4	$F_3$ Group Contribution Technique	310
A.10	Ideal Gas Heat Capacity	312
A.10.1	Group Contribution Technique	312
A.10.2	Equation Oriented Technique	313
A.11	Liquid Heat Capacity	315
A.11.1	Equation Oriented Technique	316
A.12	Glass Transition Temperature	316
A.12.1	Group Contribution Technique	317
A.13	Gas Permeability	321
A.13.1	Group Contribution Technique	321
A.14	Volume Resistivity	324
A.14.1	Equation Oriented Technique	325
A.15	Molar Polarization	325
A.15.1	Group Contribution Technique	326
A.16	Molar Volume	326
A.16.1	Group Contribution Technique	329
A.17	Molecular Weight	335
A.18	Polymer Thermal Conductivity	335
A.18.1	Equation Oriented Technique	338
A.19	Solid Molar Heat Capacity	339
A.19.1	Group Contribution Technique	340
A.20	Rao Function	343

A.20.1 Group Contribution Technique . . . . .	343
A.21 Solubility Parameters . . . . .	346
A.21.1 $\delta_p$ Group Contribution Technique . . . . .	347
A.21.2 $\delta_h$ Group Contribution Technique . . . . .	348
A.22 Drug Design . . . . .	350
A.22.1 $\sigma_M$ . . . . .	352
A.22.2 $\pi$ . . . . .	352
<b>B Estimation Procedures</b>	<b>355</b>
B.1 $P_{vp}$ Estimation Procedures . . . . .	355
B.1.1 Interactive Estimation Procedure . . . . .	355
B.1.2 Automatic Estimation Procedure . . . . .	356
B.2 $H_v$ Estimation Procedures . . . . .	359
B.2.1 Interactive Estimation Procedure . . . . .	359
B.2.2 Automatic Estimation Procedure . . . . .	359
B.3 $C_{pL}$ Estimation Procedures . . . . .	361
B.3.1 Interactive Design Procedure . . . . .	361
B.3.2 Automatic Design Procedure . . . . .	363
<b>C Physical Property Ranges</b>	<b>367</b>
C.1 Critical Temperature . . . . .	367
C.2 Critical Pressure . . . . .	369
C.3 Reduced Boiling Point . . . . .	369
C.4 Normal Boiling Point . . . . .	369
<b>D Monotonicity Identification</b>	<b>373</b>
D.1 Acentric Factor . . . . .	374
D.2 Vapor Pressure . . . . .	377
D.2.1 $P_{vp}$ Monotonicity . . . . .	378
D.2.2 $h$ Monotonicity . . . . .	380
D.2.3 $G$ Monotonicity . . . . .	380

D.2.4 <i>k</i> Monotonicity . . . . .	381
<b>E Factor Analysis</b>	<b>382</b>
E.1 Principal Component Analysis . . . . .	383
E.2 Factor Analysis . . . . .	384
E.3 Factor Analytic Studies . . . . .	384
E.3.1 Cramer . . . . .	385
E.3.2 Klinewicz . . . . .	387
E.3.3 Joback . . . . .	389
E.4 Dimensional Reduction . . . . .	391
E.5 Group Contributions . . . . .	392

# List of Figures

2.1	General Structure of 9-Anilinoacridine	12
2.2	Critical Solution Temperatures of Three Liquid Mixtures	26
2.3	Solubility Parameter Scale used in Coatings Design	35
2.4	Initial Penicillin for Modification by Drug Design	39
2.5	Input Data for Heuristic DENDRAL	44
2.6	General Design of Heuristic DENDRAL's Five Major Sections	45
4.1	Selectivity of <b>S</b> for the Separation of <b>B</b> from <b>A</b>	59
4.2	Basic Refrigeration System	63
4.3	Hypothetical Refrigeration Cycle	64
6.1	Molecules Demonstrating Structural Constraint 3	87
6.2	Example Chemical Constraint	91
6.3	Automatic Design Algorithm Structure	93
6.4	Automatic Design Meta-Algorithm	109
7.1	Graphical Representation of Constraints in a Physical Property Space	135
7.2	Graphical Representation of $T_b$ and $T_m$ Contributions	136
7.3	An Interactively Designed Molecule: Chloropropane	137
7.4	Simultaneous Design in Multiple Physical Property Spaces	140
7.5	$P_c^*-T_b^*$ Design Space for Acentric Factor Constraint	143
7.6	Two Dimensional Space Showing Four Constraints	147
7.7	Effect of Modifying $\Delta H_v$ Constraint on the Feasible Region	149
7.8	Expanded Feasible Region Formed by Relaxing Constraints by 20%	150

7.9	Interactive Design Implementation Display . . . . .	158
7.10	Temporary Group Vector . . . . .	159
7.11	Restriction of Displayed Groups . . . . .	160
7.12	Choosing the Methyl Group Vector . . . . .	162
7.13	Angle Restriction on Group Vectors . . . . .	163
7.14	Choosing a Second Terminator . . . . .	164
11.1	Basic Refrigeration System . . . . .	193
11.2	Hypothetical Refrigeration Cycle . . . . .	194
11.3	Refrigerant Design: Interactive Design Space . . . . .	208
11.4	Refrigerant Design: Freon 12 . . . . .	209
11.5	Refrigerant Design: Example Molecule . . . . .	210
11.6	Chlorine Group Vector . . . . .	214
11.7	Candidate Groups for Chlorine Replacement . . . . .	215
13.1	Schematic of Solvent Extraction . . . . .	240
13.2	Example $\delta_p$ vs. $\delta_H$ Parameter Space . . . . .	244
13.3	Example $\delta_p$ vs. $\delta_H$ Parameter Space . . . . .	245
13.4	$\delta_p$ vs. $\delta_H$ Design Space . . . . .	246
13.5	Solubility Parameter Design Space . . . . .	248
13.6	Example Solvent . . . . .	249
13.7	Solvent Mixtures . . . . .	251
13.8	Interactive Design of Solvent Mixtures . . . . .	253
14.1	Existing Antiallergic Sodium Cromoglycate . . . . .	262
14.2	“Lead” Compound for Drug Design . . . . .	263
14.3	$\sigma_M$ vs. $\pi$ Drug Design Space . . . . .	268
14.4	Two High Activity Meta-Substituents . . . . .	269
D.1	Acentric Factor as a Function of $T_{b_r}$ and $P_c$ . . . . .	375

# List of Tables

1.1	Physical Property Dependence of Chemical Products . . . . .	2
2.1	Some Estimatable Physical Properties . . . . .	10
2.2	Specific Substructure Descriptors . . . . .	13
2.3	18 Descriptors for Antitumor Activity Classification . . . . .	14
2.4	$\chi$ and $T_b$ Data for Alkanes . . . . .	17
2.5	Alkane Group Occurrences . . . . .	18
2.6	$\Delta G_{f,298}^{\circ}$ Group Contribution Estimation Errors . . . . .	19
2.7	$T_c$ Equation Oriented Estimation Technique Errors . . . . .	21
2.8	Properties Available from the Potential Energy Function . . . . .	23
2.9	Godfrey's Standard Solvents and Miscibility Number . . . . .	24
2.10	Ordering of Deviations from Raoult's Law . . . . .	28
2.11	Groups Used in Solvent Design . . . . .	29
2.12	Groups used in Polymer Design . . . . .	30
2.13	Group Contributions used in Polymer Design . . . . .	31
2.14	12 Satisfactory Candidate Polymers . . . . .	32
2.15	Military Specifications for Air Craft Coatings . . . . .	34
2.16	Erosive Aircraft Fluid's Solubility Parameters . . . . .	35
2.17	Input Data for QSAR Analysis . . . . .	39
2.18	QSAR Relationships Regressed from Input Data . . . . .	40
2.19	Estimated Activity for QSAR Analysis . . . . .	41
2.20	Heuristic DENDRAL's Identification Rules . . . . .	47
4.1	Important Physical Properties in Solvent Design . . . . .	58

5.1	An Example Estimation Procedure . . . . .	77
5.2	Group Sets for GCT-1 and GCT-2 . . . . .	77
6.1	Initial Set of Groups . . . . .	81
6.2	Example Candidate Molecules . . . . .	83
6.3	Combinatorics of Group Selection . . . . .	94
6.4	Example Meta-Groups . . . . .	96
6.5	$T_b$ Group Contributions for Meta-Group 2 . . . . .	97
6.6	Meta-Contributions . . . . .	99
6.7	$T_b$ Values for Four Meta-Molecules . . . . .	105
6.8	$T_b$ Values for 13 Meta-Molecules . . . . .	107
6.9	Hypothetical Meta-Groups from Partitioning 4 Groups into 2 Clusters .	110
6.10	Partitioning 19 Groups into N Clusters . . . . .	111
6.11	Maximum Number of Meta-Molecules . . . . .	115
6.12	Maximum Number of Meta-Molecules . . . . .	116
6.13	Meta-Molecules Formed from 3 Meta-Groups . . . . .	116
6.14	Maximum Number of Meta-Molecules Needing Testing . . . . .	118
6.15	Maximum Number of Meta-Molecules Needing Testing . . . . .	119
6.16	Pruning Results for $k=44$ , $n=3$ Automatic Design . . . . .	121
6.17	Pruning Results for $k=44$ , $n=5$ Automatic Design . . . . .	122
6.18	Example Pruning Percentage: 3 Occurrence . . . . .	123
6.19	Example Pruning Percentage: 5 Occurrence . . . . .	124
6.20	Average Number of Children Meta-Molecules Needing Testing . . . . .	125
6.21	Automatic Design with 10% Average Pruning . . . . .	126
6.22	Automatic Design with 10% Average Pruning . . . . .	127
6.23	Advantage of Abstraction: $MG_2$ Contains 1 Group . . . . .	129
7.1	Example of Linear Group Contribution Estimation Techniques . . . . .	131
7.2	$\Delta H_{vb}$ Group Contribution Estimation Technique . . . . .	138
7.3	Equations Relating Physical Properties to Factors . . . . .	145
8.1	Four Enumerated Molecules . . . . .	167

8.2	Combinatorics of Bond Association . . . . .	169
8.3	Proto-Molecules . . . . .	176
9.1	Disallowed Substructures . . . . .	179
9.2	Example Bond Lists . . . . .	182
9.3	Two Substructures in Fisher Projections . . . . .	182
9.4	Molecule – Substructure Matching Pair . . . . .	183
9.5	Atom and Bond Counts . . . . .	185
11.1	Current Refrigerants . . . . .	196
11.2	Physical Properties of Freon-12 . . . . .	199
11.3	Consistent Groups for Refrigerant Design . . . . .	203
11.4	Consistent Groups for Automatic Refrigerant Design . . . . .	206
11.5	Designed Refrigerants . . . . .	211
11.6	Estimated Property Values for Designed Refrigerants . . . . .	212
11.7	Literature Values for some Designed Refrigerants . . . . .	212
11.8	Refrigerant Design – Automatic Results . . . . .	218
12.1	Some Physical Properties of Polyimides . . . . .	223
12.2	Barrier Polymers . . . . .	225
12.3	Consistent Groups for Polymer Design . . . . .	228
12.4	Polymer Design – Automatic Results . . . . .	230
13.1	Acetic Acid Distribution Coefficients in Various Solvents . . . . .	238
13.2	$K_D$ for Several Homologous Series . . . . .	239
13.3	Acetic Acid and Water Solubility Parameters . . . . .	250
13.4	Liquid-Liquid Extraction Solvents . . . . .	251
13.5	Solvent Mixtures . . . . .	252
13.6	UNIFAC Groups . . . . .	254
13.7	Some UNIFAC Interaction Parameters . . . . .	254
14.1	Data for Initial 19 Analogs . . . . .	264
14.2	Comparison of Experimental and Model Activities . . . . .	265

14.3 Consistent Groups for Drug Design . . . . .	267
A.1 Estimated Physical Properties . . . . .	290
A.2 $T_b$ Group Contributions . . . . .	292
A.3 $T_{br}$ Group Contributions . . . . .	295
A.4 $P_c$ Group Contributions . . . . .	298
A.5 $\Delta H_{vb}$ Group Contributions . . . . .	303
A.6 $F_1$ Group Contributions . . . . .	307
A.7 $F_2$ Group Contributions . . . . .	309
A.8 $F_3$ Group Contributions . . . . .	311
A.9 $C_{pv}^o$ Group Contributions . . . . .	314
A.10 $Y_g$ Group Contributions . . . . .	319
A.11 $A$ and $S$ Parameters for Permachor Estimation . . . . .	322
A.12 $\pi$ Group Contributions . . . . .	323
A.13 Gas Permeability Estimates . . . . .	324
A.14 $P_{LL}$ Group Contributions . . . . .	327
A.15 Molar Volume Group Contributions . . . . .	330
A.16 Molar Volumes of Rubbery Amorphous Polymers at 25°C . . . . .	332
A.17 Molar Volumes of Glassy Amorphous Polymers at 25°C . . . . .	332
A.18 Molar Volume Group Contributions . . . . .	334
A.19 $M_w$ Group Contributions . . . . .	337
A.20 Thermal Conductivities of Amorphous Polymers . . . . .	339
A.21 $C_{ps}$ Group Contributions . . . . .	341
A.22 $U$ Group Contributions . . . . .	344
A.23 $\delta_p$ Group Contributions . . . . .	348
A.24 Example $\delta_p$ Estimation Errors . . . . .	349
A.25 $\delta_h$ Group Contributions . . . . .	350
A.26 Example $\delta_h$ Estimation Errors . . . . .	351
A.27 $\sigma_M$ Group Contributions . . . . .	353
A.28 $\pi$ Group Contributions . . . . .	354

B.1	$P_{vp}$ Estimation Procedure Errors – Interactive . . . . .	357
B.2	$P_{vp}$ Estimation Procedure Errors – Automatic . . . . .	358
B.3	$H_v$ Estimation Procedure Errors – Interactive . . . . .	360
B.4	$H_v$ Estimation Procedure Errors – Automatic . . . . .	362
B.5	$C_{pL}$ Estimation Procedure Errors – Interactive . . . . .	364
B.6	$C_{pL}$ Estimation Procedure Errors – Automatic . . . . .	366
C.1	$T_c$ High and Low Sample Values . . . . .	368
C.2	$P_c$ High and Low Sample Values . . . . .	370
C.3	$T_{b_r}$ High and Low Sample Values . . . . .	371
C.4	$T_b$ High and Low Sample Values . . . . .	372
D.1	Acentric Factor . . . . .	375
E.1	Physical Properties as Functions of BC(DEF) Factors . . . . .	386
E.2	Statistics for BC(DEF) Factors - Physical Property Relationships . . . . .	388
E.3	Physical Properties in Klincewicz's Factor Analytic Study . . . . .	388
E.4	Percentage Variance Explained by Klincewicz's Factors . . . . .	389
E.5	Klincewicz's Factor Analysis Data Summary Statistics . . . . .	390
E.6	Loadings for Klincewicz's Three Factor Model . . . . .	390
E.7	Physical Properties in Joback's Factor Analytic Study . . . . .	391
E.8	Total Variance Explained in Joback's Factor Models . . . . .	391
E.9	Physical Property - Factor Relationships . . . . .	392
E.10	Percentage Variance Explained by 2 Factors . . . . .	393

# Notation

## Physical Properties

$C_{p,a}^o, C_{p,b}^o, C_{p,c}^o, C_{p,d}^o$  ..... Coefficients of a cubic fit of ideal gas heat capacity.

$C_{p,L}$  ..... Liquid Heat Capacity.

$C_p^s$  ..... Solid heat capacity.

$C_{p,v}^o$  ..... Ideal Gas Heat Capacity at 298K.

$C^R(k, n)$  ..... the number of ways  $n$  objects can be chosen from a set of  $k$  objects ignoring the order of choice and allowing repetitions.

$F_1$  ..... Factor 1.

$F_3$  ..... Factor 3.

$\Delta G_{f,298}^o$  ..... Standard Gibbs Energy of Formation at 298K.

$\Delta H_{f,298}^o$  ..... Standard Enthalpy of Formation at 298K.

$\Delta H_m$  ..... Enthalpy of Fusion.

$\Delta H_v$  ..... Enthalpy of Vaporization.

$\Delta H_{vb}$  ..... Enthalpy of Vaporization at the normal boiling point.

$K_D$  ..... Equilibrium distribution coefficient.

$M$  ..... Molecular weight.

$m_B$  ..... Distribution coefficient of solute B.

$MR$  ..... Molar Refraction.

$n_A$  ..... Number of atoms.

$P_c$	Critical Pressure.
$pI_{50}$	log of the concentration producing a 50% inhibition.
$P_{LL}$	Molar polarizability.
$P(O_2)$	Permeability of oxygen through polymer.
$P_{vp}$	Vapor Pressure.
$R$	Volume resistivity.
$T_b$	Normal Boiling Point.
$T_{br}$	Reduced Boiling Point: $T_b/T_c$ .
$T_c$	Critical Temperature.
$T_g$	Glass transition temperature.
$T_m$	Normal Melting Point.
$U$	Rao parameter.
$V$	Solid molar volume.
$V_c$	Critical Volume.
$Y_g$	van Krevelen's glass transition temperature function.
$Z_c$	Critical compressibility.

### Foreign Symbols

$\delta_m$	Dipole Moment.
$\delta_p$	Polar solubility parameter.
$\delta_H$	Hydrogen bonding solubility parameter.
$\Delta_{i,PP}$	The contribution of group $i$ toward physical property $PP$ in a group contribution estimation technique.
$\eta_L$	Liquid viscosity.

$\pi$	Hansch's hydrophobicity parameter.
$\pi$	Salame's permachor.
$\lambda$	Solid thermal conductivity.
$\sigma$	Hammett's constant.
$\sigma_M$	Hammett's constant for meta-substituents.
$\sigma_P$	Hammett's constant for para-substituents.
$\chi$	Topological index.
$\omega$	Acentric Factor.

### Intervals

$X = [\underline{x} \quad \bar{x}]$  ..... An interval is denoted in two ways throughout the thesis. When referred to as a variable it is denoted by a capital letter. When referred to as a value it is denoted as a pair of symbols or numbers between brackets. The left symbol or number is the lower bound of the interval and the right symbol or number is the upper bound of the interval.

### Abbreviations

EOT	Equation Oriented estimation Technique.
GCT	Group Contribution estimation Technique.
IC	Integrated Circuit.
MG	Meta-Group.

# Preface

For my PhD thesis I developed two search techniques for designing molecules possessing desired physical property values. These techniques were incorporated into a molecular design methodology. Concepts from chemical engineering, computer science, and artificial intelligence were integrated into this systematic approach to molecular design. I implemented my methodology into a computer-based molecular design system. The implementation was done in LISP on a Symbolics 3650 LISP Machine.

My dissertation is divided into two volumes. Volume 1 describes the research area and research findings. I present my methodology along with four case studies showing its capabilities. Volume 2 details the implementation. A section by section description shows the system's operation and describes particular implementation issues.

# Chapter 1

## Introduction

Physical properties have major impact on the economics of many processes and the viability of many products. The refrigerant in a refrigeration cycle, the working fluid in a power cycle, and the solvent used in an azeotropic distillation all determine the physical and economic feasibility of the process. Chemical products such as artificial sweeteners, lubricants, and textiles all must exhibit specific physical properties for acceptance as a viable product. Table 1.1 lists a number of chemical products for which physical properties are important for good performance.

For my thesis research I developed and implemented a methodology capable of designing molecules possessing a set of desired physical properties. These techniques not only screen existing compounds but are able to generate new compounds. In this chapter I describe the incentive, objective, scope, and overview of my thesis work and this dissertation.

Table 1.1: Physical Property Dependence of Chemical Products

Chemical Product	Important Physical Properties	References
Polymer Membranes for Gas Separation	permeability, separation factor	[68],[85],[110]
Barrier Polymers for Food Packaging	permeability, glass transition temperature, toxicity, clarity, high modulus	[1],[36],[70],[110],[117]
Artificial Sweeteners	sweetness, toxicity, solubility, color, crystallinity	[30],[90]
Refrigerants	vapor pressure, liquid heat capacity, vapor volume, enthalpy of vaporization	[28],[94],[113]
Power Cycle Working Fluids	vapor pressure, critical temperature	[73]
Azeotropic Distillation Solvents	selectivity, azeotrope formation, recoverability	[8],[94]
Liquid Extraction Solvents	selectivity, capacity, distribution coefficient	[126],[77]
Polymeric Coatings	solubility, mechanical flexibility, fluid resistance	[91],[124],[125]
Dyes	color, substantivity, solubility, dyebath stability, pH stability, buildup, foaming, light fastness, wash fastness	[18]
Optical Disk Substrates	light transmission, birefringence, impact strength, water absorption, water permeation, thermal distortion	[64]
Reinforcing Fibers	tensile modulus, tensile strength, elongation to break, specific gravity, thermal stability	[83]

## 1.1 Objective

The objective of my thesis work was to develop and implement a methodology for the systematic design of molecules satisfying a set of physical property constraints. Considerable knowledge exists about determining physical property values from a compound's molecular structure[79,101,127]. Chemical knowledge[76] specifies molecular substructures which lead to unstable compounds. Structural knowledge[5] restricts the ways atoms are combined into molecules. My research was to examine how these sources of knowledge could be used in a synthetic manner – to design new molecules.

## 1.2 Incentive

Until recently identifying compounds possessing desired physical property values required searching through a vast number of molecules. Much of this search was conducted at the experimental level. Chemists hypothesizing a chemical, synthesizing the material, and testing for desired properties. Many of the current drugs and pesticides are the results of this experimental generate and test. Estimates are that 3000 to 5000 compounds needed to be tested to find one useful pharmaceutical and 5000 to 8000 to find one useful pesticide[129]

Estimating physical properties eliminates many tedious and wasteful syntheses[93]. However, a systematic procedure to use these estimation techniques in a synthetic manner is still needed. Advances in the fields of computer science and artificial intelligence now enable computers to represent and manipulate information in the domain of chemistry and chemical engineering. Applying techniques from these fields to the problem

of molecular design has the potential for vastly reducing the expense of identifying new chemical products.

### 1.3 Overview

I developed two search procedures based on the generate and test search paradigm[133]. The first procedure is interactive with the search being guided by the designer. The second procedure is automatic with the computer efficiently generating and testing a large number of candidate molecules.

These two search procedures were incorporated into a methodology for molecular design. The methodology consists of six parts:

**Step 1: Problem Formulation** The first step in any design is to identify the target[118]. In molecular design our target consists of a set of constraints on important physical properties. Molecular design targets are often stated in abstract terms such as: find a stronger polymer than kevlar; develop a freon replacement; find a solvent to facilitate the separation of acetic acid from water. Taking an abstract target and developing constraints on well characterized physical properties such as tensile strength, vapor pressure, and selectivity is this first step.

**Step 2: Target Transformation** For the computer to evaluate the performance of a candidate molecule it must be able to estimate physical properties. The target transformation step develops estimation procedures which enable the evaluation of the target physical property constraints. Estimation procedures are collections of esti-

mation techniques which can estimate a compound's physical property given only its molecular structure.

**Step 3: Generate and Test** The two search procedures are incorporated into the design methodology at this step. The search procedures generate, either interactively or automatically, candidate molecules which are then tested against the target constraints using the developed estimation procedures. Satisfactory candidates are retained while unsatisfactory candidates are pruned.

**Step 4: Molecule Enumeration** The result of the generate and test design procedures is a list of molecular substructures called groups. These groups can be connected in a number of ways resulting in different molecules. This step of the methodology enumerates all possible molecules which can be formed from the generated collections of groups.

**Step 5: Molecule Screening** Once the satisfactory candidate molecules have been enumerated we have complete molecular structures. At this step I apply a set of chemical rules which identify unstable substructures within molecules. Any of the designed molecules which contain these substructures are pruned away.

**Step 6: Final Evaluation** It is sometimes necessary to modify estimation techniques for use in the generate and test procedures. Often this modification is done to remove steps in the estimation techniques which require knowledge of global molecular structure. Using groups as my design basis I only know local structure during the

design. However, once the candidate molecules are enumerated and screened, global molecular structure is known and more accurate estimation techniques can be used to further prune the candidates.

I describe each step of my methodology in this dissertation. I present four case studies which show that my methodology is capable of designing new molecules satisfying a set of physical property constraints. Chapter 2 presents much of the previous work and many of the concepts I considered when developing my methodology. Chapter 3 proposes a model of the molecular design process which serves as the basis for much of the methodology. Chapters 4 through 10 describes the steps of the methodology. Finally, Chapters 11 through 14 present four detailed case studies demonstrating the methodology.

## 1.4 Scope

To evaluate candidate molecules it is necessary to estimate their physical properties. I used existing estimation techniques. Although I modified these techniques, it was not part of my thesis research to develop new estimation techniques. However, for certain physical properties development of group contribution estimation techniques was needed.

The final results generated by my procedure are molecular structures. Although these structures have undergone some screening to check for chemical stability, identifying if and how these structures can be synthesized was not part of my research.

A great deal of interest is currently being addressed at estimating the physical prop-

erties of enzymes, superconductors, and other compounds in which the 3-dimensional structure is extremely important. The physical properties I concerned myself with in the thesis did not involve these.

# Chapter 2

## Previous Work

My molecular design methodology codified many techniques and concepts previously used in:

1. Estimating physical properties.
2. Selecting chemical products.
3. Designing chemical products.
4. AI's generate and test search paradigm.
5. Interval arithmetic.

This chapter discusses previous work done in these areas.

### 2.1 Estimating Physical Properties

When designing chemical equipment, analyzing experimental results, or identifying chemical products, physical property values are needed. Too frequently experimental values are unknown[101]. Physical property estimation techniques were developed to satisfy this need.

Estimation techniques are available for thermodynamic properties[101], environmental properties[79], and polymer properties[127]. Estimating biological activity is the major thrust of drug design[39,53,84]. Table 2.1 presents a brief list of physical properties whose values can be estimated.

Many approaches are taken to relate molecular structure to physical properties. I classify physical property estimation techniques into five categories:

1. pattern recognition
2. topological
3. group contribution
4. equation oriented
5. molecular modeling

I briefly describe each of these categories in the following sections. Each section contains a short discussion of the concepts used followed by an example.

### 2.1.1 Pattern Recognition

Pattern recognition techniques are often employed when causal relationships are not well understood. Discriminant analysis and classification are two statistical techniques used in pattern recognition. Both are multivariate techniques concerned with *separating* distinct sets of objects into classes and with *allocating* new objects to previously defined classes. Discriminant analysis develops a discriminant function which classifies new compounds based on their molecular features. In many applications the number of classes equals two: carcinogenic or noncarcinogenic; toxic or nontoxic; etc.

Discriminant analysis begins with a set of observations and a desire to separate them into two or more classes. Each observation has a set of “features” used in the

Table 2.1: Some Estimatable Physical Properties

Symbol	Property
<b>Thermodynamic Properties</b>	
$T_b$	Normal Boiling Point
$T_m$	Normal Melting Point
$T_c, P_c, V_c$	Critical Properties
$\omega$	Acentric Factor
$P_{vp}$	Vapor Pressure
$\Delta H_{vb}$	Enthalpy of Vaporization at $T_b$
$\Delta H_v$	Enthalpy of Vaporization as a Function of Temperature
$\Delta H_{f,298}^\circ$	Enthalpy of Formation at Standard Conditions
$\Delta G_{f,298}^\circ$	Gibbs Energy of Formation at Standard Conditions
$C_p^\circ$	Ideal Gas Heat Capacity
$C_{pL}$	Liquid Heat Capacity
<b>Transport Properties</b>	
$\eta$	Viscosity
$\lambda$	Thermal Conductivity
$D_{AB}$	Binary Diffusion Coefficient
$\sigma$	Surface Tension
<b>Environmental Properties</b>	
	Adsorption Coefficient for Soils and Sediments
	Bioconcentration Factor in Aquatic Organisms
	Rate of Aqueous Photolysis
	Volatilization from Water
	Volatilization from Soil
$T_f$	Flash Points of Pure Substances
<b>Drug Properties</b>	
$\sigma$	Hammett Constant
$\pi$	Hansch Hydrophobicity Parameter
$E_S$	Taft Steric Factor

Table 2.1 Continued: Some Estimatable Physical Properties

Symbol	Property
<b>Polymer Properties</b>	
$T_g$	Glass Transition Temperature
$T_m$	Crystalline Melting Temperature
$\alpha$	Thermal Expansion Coefficient
$\epsilon$	Dielectric Constant
$\chi$	Magnetic Susceptibility
$B$	Specific Bulk Modulus

classification. The objective of the analysis is to develop a discriminant function which given the features of an observation correctly classifies the compound. A subset of compounds, called the “training set”, is used to develop the discriminant function. Once the discriminant function is formed it is used for classifying new observations.

Developing a discriminant function begins by separating our training observations into separate classes. For two classes,  $\pi_1$  and  $\pi_2$ , we obtain two collections of  $p$ -dimensional feature vectors:  $X_1$  and  $X_2$ . A discriminant function is then developed which optimally classifies  $X_1$  into  $\pi_1$  and  $X_2$  into  $\pi_2$ . We assume this function is linear of the form:

$$y = \sum_{j=1}^n l_j x_j \quad (2.1)$$

where  $l_j$  is the loading of each feature and  $y$  is a new variable called a *discriminant variable*. The analysis thus transforms a multivariate observation on  $\mathbf{X}$  into a univariate observation on  $y$ .

This univariate distribution on  $y$  is then used to develop an allocation rule. Taking

$$m = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

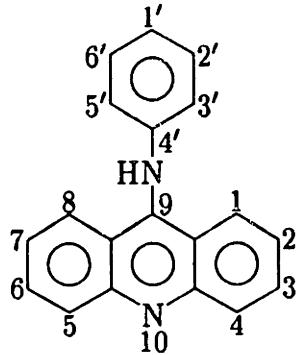


Figure 2.1: General Structure of 9-Anilinoacridine

we form the allocation rule:

If  $y_j \geq m$

Then Allocate observation  $j$  having feature vector  $X_j$  to class  $\pi_1$ .

Else Allocate observation  $j$  having feature vector  $X_j$  to class  $\pi_2$ .

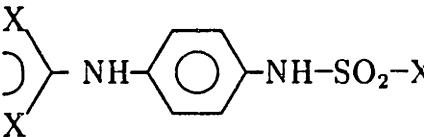
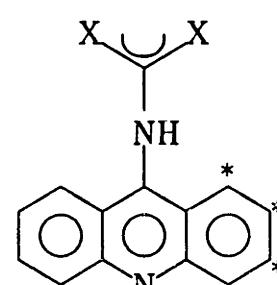
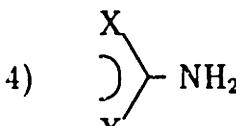
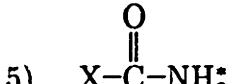
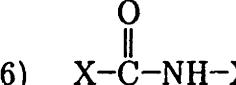
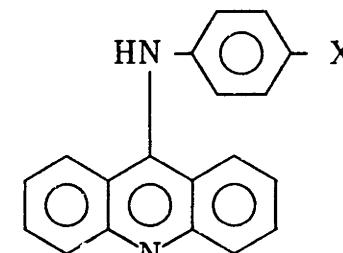
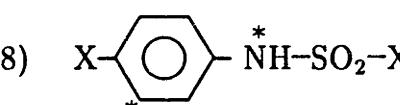
An important part of determining the classification function is to identify which features are important for classification. Additionally determining the minimum number of features facilitates theoretical interpretation of the results.

Henry, et.al.[55] performed a pattern recognition study of the antitumor activity of a set of 9-anilinoacridines. The general structure of the set of molecules is shown in Figure 2.1. Four classes of descriptors were used to separate the compounds into active and inactive classes. The four classes were:

1. Fragment descriptors - various atom, bond, and ring counts and molecular weight.
2. Topological descriptors: various  $\chi$  indices.
3. Substructure environment descriptors: the presence or absence of specific molecular substructures. Table 2.2 lists these substructures.

Table 2.2: Specific Substructure Descriptors

---

1) $X-NH_2$	9) 
2) $X-CH_2-CH_3$	
3) $X-(CH_2)_3-X$	10) 
4) 	
5) 	
6) 	11) 
7) $X-NH-\text{C}_6\text{H}_4-NH-X$	
8) 	

---

\* denotes an unspecified substituent.

4. Physicochemical property descriptors: Bondi's molecular volume[11], molar refractivity, the del-Re  $\sigma$  electronic charges at various positions on the aniline anacridine rings[25], and  $\log P$ .

A set of 18 descriptors were obtained which could correctly classify 94% of the compounds in the training set of 213. 97% of the active compounds and 85% of the inactive compounds were correctly classified. These descriptors appear in Table 2.3.

The weight vector that was obtained from the training set was applied to a prediction set of 119 compounds that were not included in the original analysis. The prediction set results had an accuracy greater than 73% indicating the usefulness of

Table 2.3: 18 Descriptors for Antitumor Activity Classification

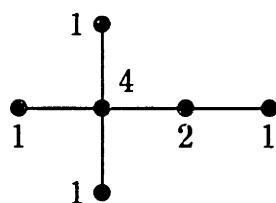
	Descriptor	Coefficient	Variance
1)	# of S atoms	0.169	0.012
2)	# of rings	-0.172	0.002
3)	Average # of paths per atom	0.326	0.004
4)	Molar Refractivity	-0.213	0.010
5)	Substructure 1	-0.141	0.011
6)	Substructure 2	0.033	0.044
7)	Substructure 3	0.090	0.029
8)	Substructure 4	0.457	0.003
9)	Substructure 5	0.083	0.027
10)	Substructure 6	-0.129	0.017
11)	Substructure 7	0.224	0.003
12)	Substructure 8	0.086	0.061
13)	Substructure 9	-0.323	0.002
14)	Substructure 10	-0.241	0.017
15)	Substructure 11	-0.128	0.004
16)	Charge Position 2	-0.142	0.016
17)	Charge Position 3	0.153	0.001
18)	Charge Position 2'	-0.333	0.018
	Constant	0.361	0.003

pattern recognition in screening active compounds.

### 2.1.2 Topological Techniques

Topological techniques ignore the actual three-dimensional shape of a molecule, the nature and lengths of the chemical bonds connecting its atoms, the angles between the bonds, and sometimes even atom types[107]. Only the number of atoms and their interconnections are considered. This information is reduced to an *index*. Some of the proposed indices are the Wiener Path Number[132], Altenburg Polynomial Index[2], Gordon and Scantlebury Index[46], Hosoya's Z Index[59], and Randić's Branching Index[96]. Randić's index has been formalized and extended by Kier and Hall[65]. I use the extension of Randić's index by Kier and Hall to demonstrate this class of estimation techniques.

Calculation of the index begins by drawing a hydrogen suppressed molecular graph. The graph for 2,2-dimethylbutane is:



The numbers correspond to a valence,  $\delta$ , assigned to each carbon atom equal to the number of C-C bonds in which the atom participates. Using the Randić algorithm the contribution of each bond toward the index is:

$$(\delta_i \delta_j)^{-0.5} \quad (2.2)$$

The total of the contributions for each bond equals the molecular connectivity index,

$\chi$ . For 2,2-dimethylbutane:

$$\chi = \frac{1}{\sqrt{1 \cdot 4}} + \frac{1}{\sqrt{1 \cdot 4}} + \frac{1}{\sqrt{1 \cdot 4}} + \frac{1}{\sqrt{4 \cdot 2}} + \frac{1}{\sqrt{2 \cdot 1}} = 2.5607$$

Kier and Hall extended the algorithm to include heteroatoms[65].

Table 2.4 shows  $\chi$  and  $T_b$  values for 39 alkanes. Linear regression of the data resulted in the equation

$$T_b = 67.5\chi - 130.0 \quad (2.3)$$

having an  $r^2$  value of 0.977.

### 2.1.3 Group Contribution Techniques

Group contribution techniques make the assumption that each fragment of a molecule contributes a certain amount to the value of its physical property. The further assumption is made that this contribution is dependent only upon the local conditions, the atom itself, its neighbors, and sometimes its neighbors neighbors.

Developing group contribution techniques begins by collecting a set of molecules having known values for the property to be estimated. A set of groups is chosen which can represent the molecules. The occurrence of these groups in each molecule is recorded.

Taking data for the Gibbs energy of formation for the alkanes we obtain the following four groups:



Table 2.5 shows the group occurrences for 15 alkanes.

Table 2.4:  $\chi$  and  $T_b$  Data for Alkanes

Alkane	$\chi$	$T_b$
Ethane	1.00000	-88.630
Propane	1.41421	-42.070
n-Butane	1.91421	-0.500
2-Methylpropane	1.73205	-11.730
n-Pentane	2.41421	36.074
2-Methylbutane	2.27005	27.852
2,2-Dimethylpropane	2.00000	9.503
n-Hexane	2.91421	68.740
2-Methylpentane	2.77005	60.271
3-Methylpentane	2.80806	63.282
2,2-Dimethylbutane	2.56066	49.741
2,3-Dimethylbutane	2.64273	57.988
n-Heptane	3.41421	98.427
2-Methylhexane	3.27005	90.052
3-Methylhexane	3.30806	91.850
3-Ethylpentane	3.34606	93.475
2,2-Dimethylpentane	3.06066	79.197
2,3-Dimethylpentane	3.18073	89.784
2,4-Dimethylpentane	3.12589	80.500
3,3-Dimethylpentane	3.12132	86.064
2,2,3-Trimethylbutane	2.94337	80.882
n-Octane	3.91421	125.665
2-Methylheptane	3.77005	117.647
3-Methylheptane	3.80806	118.925
4-Methylheptane	3.80806	117.709
3-Ethylhexane	3.84606	118.534
2,2-Dimethylhexane	3.56066	106.840
2,3-Dimethylhexane	3.68073	115.607
2,4-Dimethylhexane	3.66390	109.429
2,5-Dimethylhexane	3.62589	109.103
3,3-Dimethylhexane	3.62132	111.969
3,4-Dimethylhexane	3.71784	117.725
2-Methyl-3-Ethylpentane	3.71784	115.650
3-Methyl-3-Ethylpentane	3.68198	118.259
2,2,3-Trimethylpentane	3.48138	109.841
2,2,4-Trimethylpentane	3.41650	99.238
2,3,3-Trimethylpentane	3.50403	114.760
2,3,4-Trimethylpentane	3.55341	113.467
2,2,3,3-Tetramethylbutane	3.25000	106.470

Table 2.5: Alkane Group Occurrences

Alkane	Formula	-CH <sub>3</sub>	-CH <sub>2</sub> -	>CH-	>C<
ethane	C <sub>2</sub> H <sub>6</sub>	2	0	0	0
propane	C <sub>3</sub> H <sub>8</sub>	2	1	0	0
n-butane	C <sub>4</sub> H <sub>10</sub>	2	2	0	0
isobutane	C <sub>4</sub> H <sub>10</sub>	3	0	1	0
n-pentane	C <sub>5</sub> H <sub>12</sub>	2	3	0	0
2,2-dimethylpropane	C <sub>5</sub> H <sub>12</sub>	4	0	0	1
n-hexane	C <sub>6</sub> H <sub>14</sub>	2	4	0	0
2-methyl pentane	C <sub>6</sub> H <sub>14</sub>	3	2	1	0
3-methyl pentane	C <sub>6</sub> H <sub>14</sub>	3	2	1	0
2,2-dimethyl butane	C <sub>6</sub> H <sub>14</sub>	4	1	0	1
2,3-dimethyl butane	C <sub>6</sub> H <sub>14</sub>	4	0	2	0
n-heptane	C <sub>7</sub> H <sub>16</sub>	2	5	0	0
2-methylhexane	C <sub>7</sub> H <sub>16</sub>	3	3	1	0
3-methylhexane	C <sub>7</sub> H <sub>16</sub>	3	3	1	0
2,2-dimethylpentane	C <sub>7</sub> H <sub>16</sub>	4	2	0	1

For  $n$  compounds our data consists of an  $(n \times 1)$  vector of physical properties,  $\mathbf{P}$ , and an  $(n \times 4)$  matrix of group occurrences,  $\mathbf{G}$ . Assuming a linear relationship between the occurrences of the groups and the physical property we formulate the problem as finding the vector of 4 group contributions,  $\Delta$ , such that

$$\mathbf{G} \Delta = \mathbf{P} \quad (2.4)$$

Equation 2.4 is overdetermined and is solved using a regression technique. The most commonly used regression is least squares giving

$$\Delta = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{P} \quad (2.5)$$

Data from 54 alkanes were regressed onto our four groups. The resulting model of the regression is

$$\Delta G_{f,298}^{\circ} = -4.48n_{(-CH_3)} + 2.08n_{(-CH_2-)} + 8.70n_{(>CH-)} + 15.60n_{(>C<)} \quad (2.6)$$

Table 2.6:  $\Delta G_{f,298}^\circ$  Group Contribution Estimation Errors

	Compound	Formula	Literature	Error
1)	isobutane	$C_4H_{10}$	-4.99	0.25
2)	n-pentane	$C_5H_{12}$	-2.00	-0.72
3)	2,2-dimethylpropane	$C_5H_{12}$	-3.64	1.32
4)	n-hexane	$C_6H_{14}$	-0.06	-0.58
5)	2,3-dimethyl butane	$C_6H_{14}$	-0.98	0.46
6)	n-heptane	$C_7H_{16}$	1.91	-0.47
7)	2-methylhexane	$C_7H_{16}$	0.77	0.73
8)	3-methylhexane	$C_7H_{16}$	1.10	0.40
9)	3,3-dimethylpentane	$C_7H_{16}$	0.63	1.21
10)	3-ethylpentane	$C_7H_{16}$	2.63	-1.13
11)	2,2,3-trimethylbutan	$C_7H_{16}$	1.02	0.88
12)	n-octane	$C_8H_{18}$	3.92	-0.40
13)	2,4-dimethylhexane	$C_8H_{18}$	2.80	0.84
14)	3-ethylhexane	$C_8H_{18}$	3.95	-0.37
15)	2,2,3-trimethylpentane	$C_8H_{18}$	4.09	-0.11
16)	3,3-diethylpentane	$C_9H_{20}$	8.38	-2.38
17)	2,3,3,4-tetramethylpentane	$C_9H_{20}$	8.15	-2.03
18)	n-decane	$C_{10}H_{22}$	7.94	-0.26
19)	n-undecane	$C_{11}H_{24}$	9.94	-0.18
20)	hexadecane	$C_{16}H_{34}$	20.00	0.16

Error = Estimate - Literature. Units are kJ/mol. Literature values were from [101].

The  $r^2$  value was 0.975. Table 2.6 shows the errors of the estimation for 20 compounds.

Group contribution estimation techniques can become very complex including nonlinear effects and interactions among groups. UNIFAC[41] is an example of a complex group contribution estimation technique.

#### 2.1.4 Equation Oriented Techniques

Equation oriented estimation techniques are the most widely used type of estimation technique. One physical property is related to one or more other physical properties by

theoretical or empirical models. The goal is to relate the estimated physical properties to properties more available or more easily measured.

Equation oriented estimation techniques are typically a combination of theory and empirical observation. Theoretical concepts often suggest interrelations among physical properties or between properties and state variables. These premises instigate empirical inquiry into reifying such relationships.

Klincewicz[67] developed an equation oriented estimation technique for the critical temperature.  $T_c$  is related to  $T_b$  and  $Mw$  by

$$T_c = 50.2 - 0.16Mw + 1.41T_b \quad (2.7)$$

Table 2.7 shows the errors associated with estimating 20 compounds. Errors of between 1 to 2 percent were reported[102]. Using Equation 2.7 I estimated  $T_c$  for 406 organic compounds obtaining an average absolute error of 18.7 K giving an average absolute percent error of 3.1%.

### 2.1.5 Molecular Modeling Techniques

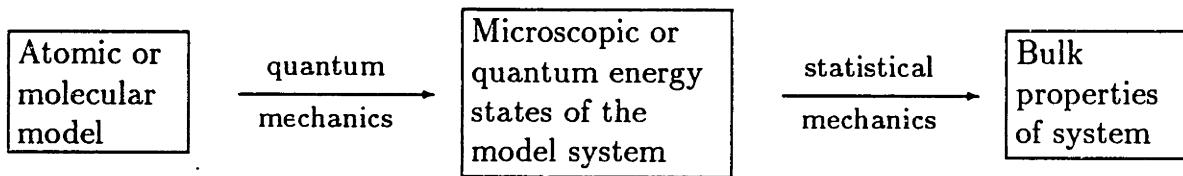
The ultimate objective of a molecular approach to physical properties is to calculate the macroscopic properties of matter from first principles. The term “first principles” refers to quantum theory and to statistical mechanics. The mathematical solution of the formulas of quantum mechanics provides us with the microscopic or molecular energy values, any one of which a system may assume at a given instant of time. With the microscopic energy levels available, the methods of statistical mechanics can be applied to give the observable or bulk properties of the molecular system.

Table 2.7:  $T_c$  Equation Oriented Estimation Technique Errors

Compound	Formula	$T_c$	Error	% Error
Dichlorodifluoromethane	$\text{CCl}_2\text{F}_2$	385.0	-11.2	-2.9
Carbon Tetrafluoride	$\text{CF}_4$	227.6	13.1	5.8
Methyl Mercaptan	$\text{CH}_3\text{S}$	470.0	-34.0	-7.2
Ketene	$\text{C}_2\text{H}_2\text{O}$	380.0	-9.4	-2.5
Acetonitrile	$\text{C}_2\text{H}_3\text{N}$	545.5	-1.6	-0.3
Acetic Acid	$\text{C}_2\text{H}_4\text{O}_2$	592.7	-0.7	-0.1
Ethane	$\text{C}_2\text{H}_6$	305.4	0.3	0.1
Glycerol	$\text{C}_3\text{H}_8\text{O}_3$	726.0	103.3	14.2
Methyl Ethyl Sulfide	$\text{C}_3\text{H}_8\text{S}$	533.0	-15.9	-3.0
Vinylacetylene	$\text{C}_4\text{H}_4$	455.0	-21.0	-4.6
Thiophene	$\text{C}_4\text{H}_4\text{S}$	579.4	-39.0	-6.7
2-Butyne	$\text{C}_4\text{H}_6$	488.7	-24.0	-4.9
n-Propyl Formate	$\text{C}_4\text{H}_8\text{O}_2$	538.0	-2.6	-0.5
n-Valeric Acid	$\text{C}_5\text{H}_{10}\text{O}_2$	651.0	30.8	4.7
Iodobenzene	$\text{C}_6\text{H}_5\text{I}$	721.0	-52.6	-7.3
Triethylamine	$\text{C}_6\text{H}_{15}\text{N}$	535.6	9.5	1.8
o-Cresol	$\text{C}_7\text{H}_8\text{O}$	697.6	-10.2	-1.5
Isobutylcyclohexane	$\text{C}_{10}\text{H}_{20}$	659.0	-4.5	-0.7
Hexadecane	$\text{C}_{16}\text{H}_{34}$	722.0	81.6	11.3

Error = Estimate - Literature. Units are K. Literature values are from [101].

The procedure is summarized by the following two steps[100]:



The objective of statistical mechanics is to show how the properties of matter in bulk (macroscopic properties) can be calculated from the properties of individual molecules (positions, molecular geometry, interatomic and intermolecular forces, etc.). Quantum mechanics alone cannot supply these macroscopic properties because it deals with the detailed dynamics of the particles.

The whole field of molecular mechanics, conformational energy calculations, and molecular dynamics simulations rests on the fact that the potential energy of a molecule or assembly of molecules can be written as an analytically simple sum of terms, involving internal coordinates of the molecule (i.e., bond lengths, bond angles, torsion angles, and interatomic distances)[48]. This potential energy function is a fundamental physical quantity, implicitly containing essentially all conformational properties of the molecular system of interest (with the exception of quantum properties). Some of these properties are presented in Table 2.8.

Molecular dynamics can be used to compute the dynamics of the molecular system. The first step is specifying the potential energy expression. Initial coordinates and velocities are specified for each of the atoms in the system. Once the initial conditions are specified, the equations of motion:

$$-\frac{\partial V}{\partial \mathbf{r}_i} = \mathbf{F} = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \quad (2.8)$$

Table 2.8: Properties Available from the Potential Energy Function

$V$	Energy Monte Carlo solvent effects
$\partial V/\partial x = 0$	Minimum-energy conformation
$\partial^2 V/\partial x^2$	Vibrational spectra Normal modes Entropy and free energy
$-\partial V/\partial x = \mathbf{F} = m\mathbf{a}$	Dynamic trajectory Conformational fluctuations

are written for each atom. The equations are then integrated forward in time to compute the trajectories of each atom.

## 2.2 Selecting Chemical Products

Selecting existing compounds for use as chemical products is a two step procedure. The first step involves identifying those physical properties which are important to the performance of the chemical product and their values which give optimal behavior. The second step is searching a data base for existing compounds possessing these physical property values.

Identifying important physical properties is not a trivial task. Thermodynamic, transport, environmental, and economic factors must be accounted for in making decisions. The procedures used for solvent selection exemplify the multiple ways in which the problem can be formulated. Three solvent selection procedures are discussed.

Table 2.9: Godfrey's Standard Solvents and Miscibility Number

1	Glycerol	17	<i>p</i> -Dioxane
2	Ethylene Glycol	18	3-Pentanone
3	1,4-Butanediol	19	1,1,2,2-Tetrachloroethane
4	2,2'-Thiodiethanol	20	1,2-Dichloroethane
5	Diethylene Glycol	21	Chlorobenzene
6	Triethylene Glycol	22	1,2-Dibromobutane
7	Tetraethylene Glycol	23	1-Bromobutane
8	Methoxyacetic Acid	24	1-Bromo-3-methylbutane
9	Dimethyl Sulfoxide	25	<i>sec</i> -Amylbenzene
10	N-Formylmorpholine	26	4-Vinylcyclohexane
11	Furfuryl Alcohol	27	1-Methylcyclohexene
12	2-(2-Methoxyethoxy) Ethanol	28	Cyclohexane
13	2-Methoxyethanol	29	Heptane
14	2-Ethoxyethanol	30	Tetradecane
15	2-(2-Butoxyethoxy) Ethanol	31	Petrolatum
16	2-Butoxyethanol		

### 2.2.1 Godfrey

Godfrey[44] addressed the problem of identifying whether or not two liquids were miscible. He selected 31 typical solvents which spanned the range from strongly lipophilic to strongly lipophobic. To each of these solvents he assigned a *Miscibility Number*.

Table 2.9 shows the 31 solvents with their assigned miscibility numbers.

Godfrey developed the following three rules to determine miscibility of two liquids:

1. All pairs of standard solvents whose miscibility numbers differ by 15 or less are miscible in all proportions at 25°C.
2. Each pair whose miscibility number difference is 16 has a critical solution temperature between 25° and 75°C.
3. A difference of 17 or more corresponds to immiscibility, or to a critical solution

temperature above 75°C.

To classify a new liquid in the miscibility number system, the chemist determines a cut-off in miscibility or immiscibility with the standard solvents. 15 is then added or subtracted to the miscibility number of the standard to give the new solvent's number. In this manner Godfrey determined the miscibility number for over 400 organic liquids.

### **2.2.2 Francis**

Francis[38] developed a solvent selection procedure emphasizing selectivity for hydrocarbons. He investigated the applicability of critical solution theory as a measure of affinity. The temperature at which a liquid mixture separates into its components is its critical solution temperature.

Figure 2.2 shows the solubility behavior of three hydrocarbons in aniline. The three hydrocarbons boil almost at the same temperature, so they can not be readily separated by fractional distillation. In solvent extraction the hydrocarbons form approximately ideal mixtures, so that the amounts extracted by a solvent are nearly proportional to the separate solubilities. Since the three curves have almost exactly the same shape except for vertical displacement, the solubilities are related simply to the critical solution temperature.

### **2.2.3 Berg**

Berg[8] developed a classification scheme for hydrogen bonding in azeotropic distillation solvents. Five classes of solvents were formed ranging from highly hydrogen bonding

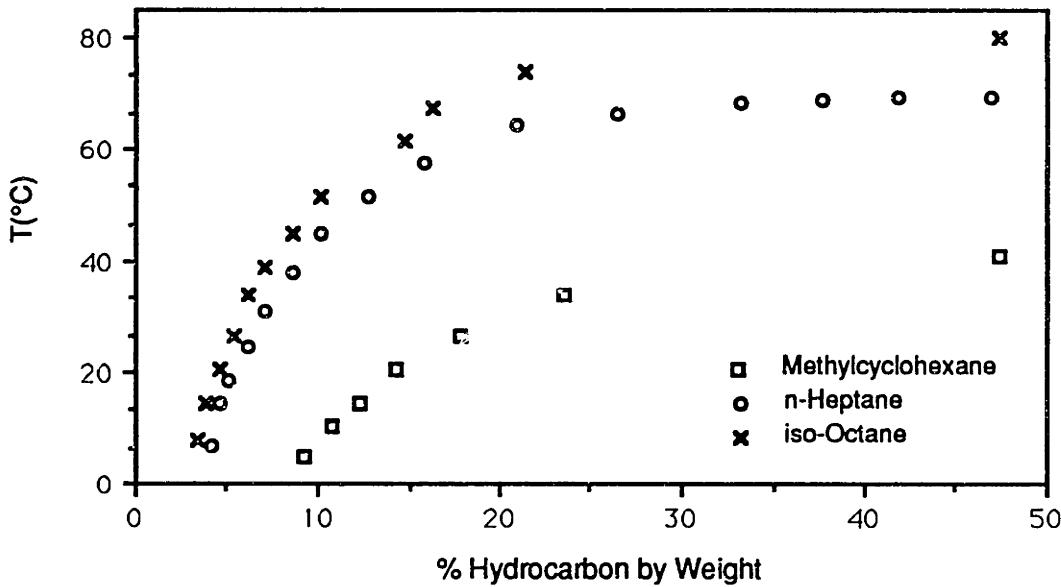


Figure 2.2: Critical Solution Temperatures of Three Liquid Mixtures

substances to non-hydrogen bonding substances. The classes are listed here:

- **Class I:** Liquids capable of forming three dimensional networks of strong hydrogen bonds – e.g.; water, glycol, glycerol, amino alcohols, hydroxylamine, hydroxy-acids, polyphenols, amides, etc.
- **Class II:** Other liquids composed of molecules containing both active hydrogen atoms and donor atoms (oxygen, nitrogen, and fluorine) – e.g.; alcohols, acids, phenols, primary and secondary amines, nitro compounds with alpha-hydrogen atoms, nitriles with alpha-hydrogen atoms, ammonia, hydrazine, hydrogen fluoride, hydrogen cyanide, etc.
- **Class III:** Liquids composed of molecules containing donor atoms but no active hydrogen atoms – e.g.; ethers, ketones, aldehydes, esters, tertiary amines, nitro

compounds, etc.

- **Class IV:** Liquids composed of molecules containing active hydrogen atoms but no donor atoms. These are molecules having two or three chlorine atoms on the same carbon atoms as a hydrogen atom, or one chlorine on the same carbon atom and one or more chlorine atoms on adjacent carbon atoms – e.g.;  $\text{CHCl}_3$ ,  $\text{CH}_2\text{Cl}_2$ ,  $\text{CH}_3\text{CHCl}_2$ , etc.
- **Class V:** All other liquids – i.e., liquids having no hydrogen bond forming capabilities – e.g.; hydrocarbons,  $\text{CS}_2$ , sulfides, mercaptans, halohydrocarbons, etc.

Table 2.10 shows the ordering of deviations from Raoult's law. Berg proposed this ranking to assist identifying solvents for use in azeotropic distillation. To form an azeotrope with a compound from Class I Table 2.10 suggests examining compounds in Class V.

## 2.3 Designing Chemical Products

Designing new chemical products is a two step approach similar to that used in selecting chemical products. The first step is identical: identify the important physical properties and their values. The second step is to search for compounds which possess acceptable values for these important physical properties. However, unlike selecting from existing compounds, we do not have a data base of existing compounds to search. The molecular structures of new compounds must be hypothesized.

Because physical properties must be determined for compounds which may have

Table 2.10: Ordering of Deviations from Raoult's Law

Classes			Deviations
I	+	V	Always + deviations; I + V,
II	+	V	frequently limited solubility.
III	+	IV	Always - deviations.
I	+	V	Always + deviations; I + IV,
II	+	IV	frequently limited solubility.
I	+	I	Usually + deviations; very
I	+	II	complicated situation, some
I	+	III	combinations give maximum
II	+	II	azeotropes.
II	+	III	
III	+	III	Quasi-ideal systems, always
III	+	V	+ deviations or ideal;
IV	+	IV	azeotropes, if any, will
IV	+	V	be minimum.
V	+	V	

never been synthesized, physical property estimation techniques play an important role in chemical product design. The three techniques discussed here form candidate molecules and estimate their properties using estimation techniques. Group contribution estimation techniques are used in all three cases.

### 2.3.1 Solvent Design

Gani and Brignole [13,42] used the UNIFAC[41] group contribution method to synthesize molecular structures with specific solvent properties for the separation of aromatic and paraffinic hydrocarbons by liquid-liquid extraction. Their synthesis procedure is divided into 3 steps:

1. Select the groups considered suitable building blocks for the molecular structures.
2. Combine the groups into candidate molecules according to specified combination rules.

Table 2.11: Groups Used in Solvent Design

---

$-\text{CH}_3$	$-\text{CH}_2-$	$-\text{COCH}_3$	$-\text{CH}_2\text{CN}$	$-\text{CH}_2\text{NO}_2$
$-(\text{C}_5\text{H}_4\text{N})$	$-\text{CH}_2\text{CO}-$	$>\text{CHNO}_2$	$-(\text{C}_5\text{H}_3\text{N})-$	$-\text{OH}$
$-\text{OCH}_3$	$\text{CH}_3\text{COO}-$	$-\text{CH}_2\text{COO}-$	$-\text{CH}_2\text{O}-$	

---

3. Screen the candidate molecules using UNIFAC to evaluate their usefulness for a particular solvent extraction task.

Chlorinated, olefinic, carboxylic, aldehyde, and aromatic groups were excluded from the design because of potential chemical instability or corrosion problems. Table 2.11 shows the resulting fourteen groups used in the design.

To reduce the number of group combinations to a tractable number several additional constraints were placed on the candidate solvents. The solvents should have a rather high boiling point to obtain a simple separation from the aromatic fraction and to avoid the formation of azeotropes. This constraint on the boiling point was propagated to a constraint on the minimum value for molecular weight being 100. A maximum value was chosen to be 140. Another constraint was that a candidate solvent should have at least one nonhydrogen group.

In one analysis 85 compounds were obtained after screening according to chosen groups, structural constraints, and molecular weight range. These were tested using UNIFAC for their use as a solvent for the separation of n-heptane and toluene. Fourteen compounds were found to have satisfactory solvent properties.

Table 2.12: Groups used in Polymer Design

$-\text{CH}_2-$	$-\text{CO}-$	$-\text{COO}-$
$-\text{O}-$	$-\text{CONH}-$	$-\text{CHOH}-$
	$-\text{CHCl}-$	

### 2.3.2 Polymer Design

Derringer and Markham[26] proposed a generate and test methodology for designing polymers possessing desired physical properties. They based their methodology upon Krevelen's[127] group contribution estimation techniques. A computer program was written to find viable polymer structures for meeting constraints on three physical properties:

1. density,  $D$
2. water absorption,  $W$
3. glass transition temperature,  $T_g$ .

The groups comprising the repeat unit were limited to the seven shown in Table 2.12.

Van Krevelen described estimation models for the three physical properties:

$$D = \sum M_i / \sum V_{gi} \quad \text{g/cm}^3$$

$$W = 18 \sum H_i / \sum M_i \quad \text{g H}_2\text{O/g polymer}$$

$$T_g = (\sum Y_{gi} / \sum M_i) - 273 \quad ^\circ\text{C}$$

$M_i$  is the contribution of the  $i$ th group to the gram molecular weight,  $V_{gi}$  the contribution to the molar volume,  $H_i$  the contribution to the molar water content, and  $Y_{gi}$  the contribution to the molar glass transition temperature function. The group contributions are given in Table 2.13.

Table 2.13: Group Contributions used in Polymer Design

Group	$M_i$	$V_{gi}$	$H_i$	$Y_{gi}$
$-\text{CH}_2-$	2,700	15.85	$3.3 \times 10^{-5}$	14.0
$-\text{CO}-$	27,000	13.40	0.110	28.0
$-\text{COO}-$	8,000	23.00	0.075	44.0
$-\text{O}-$	4,000	10.00	0.200	16.00
$-\text{CONH}-$	12,000	24.90	0.750	43.0
$-\text{CHOH}-$	13,000	19.15	0.750	30.0
$-\text{CHCl}-$	20,000	29.35	0.015	48.5

The methodology begins with a set of physical property constraints. These represent the design target. Candidate polymers are generated by selecting collections of groups from the seven groups shown in Table 2.12. The generate and test procedure chooses a collection of groups, estimates their physical properties, and checks these against the target. Those candidates which satisfy the tests are kept while those which fail the test are pruned away.

For the physical property target:

$$1.0 < \text{Density} < 1.5 \text{ g/cm}^3$$

$$0.0 < \text{Water absorption} < 0.18 \text{ g(H}_2\text{O)/g(polymer)}$$

$$T_g > 25^\circ\text{C}$$

the computer program generated 300 candidate polymers randomly selecting both the number of groups per structural repeat unit and the groups themselves. Only 12 of these candidates satisfied the target specifications. These are shown in Table 2.14.

Recognizing that the number of candidates which are identified may be large, Derringer and Markham devised a ranking procedure. This procedure consists of the following two steps:

Table 2.14: 12 Satisfactory Candidate Polymers

Repeat Unit	$g(H_2O)/g(\text{polymer})$	$T_g(\text{ }^\circ\text{C})$	Density (g/cm <sup>3</sup> )
$-(\text{CH}_2-\text{CHCl})-$	0.004	90	1.38
$-(\text{CH}_2-\text{COO}-\text{CH}_2)-$	0.019	79	1.32
$-(\text{CO}-\text{CH}_2)-$	0.047	433	1.44
$-(\text{CO}-\text{CH}_2-\text{O}-\text{CH}_2)-$	0.032	232	1.31
$-(\text{CHCl}-\text{CHCl}-\text{CH}_2)-$	0.005	112	1.49
$-(\text{CH}_2-\text{COO}-\text{CH}_2-\text{CO})-$	0.033	251	1.47
$-(\text{CHCl}-\text{CH}_2-\text{CHOH}-\text{O})-$	0.130	93	1.46
$-(\text{CO}-\text{CH}_2-\text{CH}_2)-$	0.035	305	1.24
$-(\text{CH}_2-\text{CO}-\text{CH}_2-\text{CONH})-$	0.156	175	1.42
$-(\text{CO}-\text{O}-\text{CH}_2)-$	0.040	308	1.48
$-(\text{CHCl}-\text{CH}_2-\text{O})-$	0.008	67	1.42
$-(\text{CH}_2-\text{CHOH}-\text{CHCl})-$	0.149	113	1.44

1. For each predicted property the property level  $Y$  is transformed to a 0-1 desirability,  $d_i$ , scale. On this scale, 0, corresponds to a property level that makes the polymer useless for the application being considered. A desirability of 1.0, on the other hand, corresponds to a property level such that no other property level would make the polymer more useful for the application. For intermediate values, the higher the  $d$ , the more desirable is the corresponding property value.

Property constraints are either one-sided or two-sided. If a property value must be greater than some value or less than some value, the constraint is one-sided. If the property value must be between certain limits, the constraint is two-valued.

For a one-sided constraint the desirability function is:

$$d_i = \begin{cases} 0 & \hat{Y}_i \leq Y_{i*} \\ \left[ \frac{\hat{Y}_i - Y_{i*}}{Y_i^* - Y_{i*}} \right]^r & Y_{i*} \leq \hat{Y}_i \leq Y_i^* \\ 1 & \hat{Y}_i \geq Y_i^* \end{cases}$$

Here  $\hat{Y}_i$  represents the predicted property value and  $Y_{i*}$  is the minimum acceptable level of property  $Y_i$ . Values below this value will have a desirability of zero.  $Y_i^*$  is the highest value of  $Y_i$  which will translate into improved utility. The desirability at  $Y_i^*$  equals 1.0 and does not increase as  $Y_i$  increases above  $Y_i^*$ . The value of  $r$  is adjustable and determines the rate of increase of  $d_i$  with  $Y_i$  between  $Y_{i*}$  and  $Y_i^*$ .

For a two-sided constraint the desirability function is:

$$d_i = \begin{cases} \left[ \frac{\hat{Y}_i - Y_{i*}}{c_i - Y_{i*}} \right]^s & Y_{i*} \leq \hat{Y}_i \leq c_i \\ \left[ \frac{\hat{Y}_i - Y_i^*}{c_i - Y_i^*} \right]^t & c_i \leq \hat{Y}_i \leq Y_i^* \\ 0 & \hat{Y}_i < Y_{i*} \text{ or } \hat{Y}_i > Y_i^* \end{cases}$$

Here  $Y_{i*}$  and  $Y_i^*$  are, respectively, the lower and upper constraints on property  $Y_i$  below and above which the desirability will equal zero. Parameter  $c_i$  is the most desirable level of the property  $Y_i$  and corresponds to a desirability of 1.0. The rate of decrease of  $d_i$  above and below  $c_i$  is determined by adjustable parameters  $t$  and  $s$ .

2. The individual desirabilities are combined into a composite desirability  $D$  by taking the geometric mean of the individual desirabilities as follows:

$$D = (d_1 d_2 d_3 \dots d_k)^{1/k} \quad (2.9)$$

where  $k$  is the number of properties.

### 2.3.3 Polymer Coatings Design

Tortorello and Kinsella[124,125] used the solubility parameter concept to design high performance coatings for aircraft. Legislative reform and a changing perspective regarding the supply and economic advantages of solvents suggested a change from the Air Force's solvent-based epoxy-polyamide primer and urethane topcoat to a water-based coating. Table 2.15 summarizes the military specifications which characterize the properties of the solvent-based urethane topcoat[124]. The requirements of the primer are similar. Any prospective replacement should display equivalent performance.

The results of a preliminary screening of the commercial marketplace for aqueous resins failed to identify a satisfactory candidate. Tortorello and Kinsella thus proceeded to design a new polymer.

The aircraft coating would have to be designed to resist five fluids. These five fluids and their solubility parameters are shown in Table 2.16. Figure 2.3 shows these fluids

Table 2.15: Military Specifications for Air Craft Coatings

---

5% Salt spray	No blistering, cracking, corrosion, or loss of adhesion after 500 hours of exposure.
100% Relative humidity	No blistering, cracking, softening, or loss of adhesion after 720 hours of exposure.
Accelerated weathering	After 500 hour exposure the coating should exhibit 60% impact flexibility, no more than 10% loss of original gloss, and no color change.
Fluid resistance	A decrease of no more than one pencil hardness unit after immersion in water (4 days, 100°F), lubricating oil (24 hours, 250°F), hydrocarbon fluid (7 days, room temperature), and hydraulic fluid (7 days, room temperature). A decrease of no more than two pencil hardness units after immersion in Skydrol 500B fluid (7 days, room temperature).
Film flexibility	No cracking, crazing, or loss of adhesion of coating when elongated 60% by impacting mandrel.
Low temperature flexibility	No cracking or loss of adhesion when bent around $\frac{3}{8}$ inch diameter cylindrical mandrel after 4 hours at -65°F.
High temperature resistance	No loss of adhesion or flexibility after 4 hours at 300°F.
60° Gloss	>90

---

Table 2.16: Erosive Aircraft Fluid's Solubility Parameters

Compound	Solubility Parameter
H5606 Hydraulic Fluid	7.0
TT-S-735 Hydrocarbon	7.5
Lubricating Oil	8.0
Skydrol 500B Lubricating Fluid	11.0
Water	23.0

on a solubility parameter scale.

Hildebrand[56] showed that for a solution process to occur the solubility parameter value of the solvent must be nearly equal to that of the solute. Conversely, incompatibility is predicted when there is a disparity between the two values. Hence, the design of novel polymers for enhanced fluid resistance can be guided by a broad distinction between the solubility parameter value of the polymer and that of the fluid. Tortorello and Kinsella noticed the large gap between the Skydrol 500B fluid and water. They sought to design a resin which would have a solubility parameter falling within the gap.

Using the group contributions of Small[115], Rheineck and Lin[103], and Fedors[33,

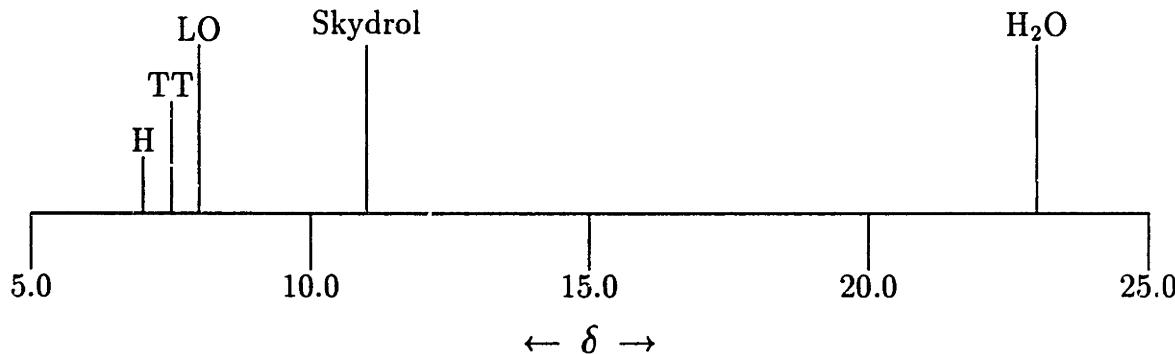


Figure 2.3: Solubility Parameter Scale used in Coatings Design

34], Tortorello and Kinsella designed a series of acrylic polymers varying in solubility parameter. The series varied in solubility parameter from 11 to 14.

Tortorello and Kinsella synthesized and tested series of urethane and acrylic polymers. The experimental results corroborated with the solubility parameter model predictions. They concluded that the solubility parameter concept was useful for identifying polymers with desired fluid resistance.

#### 2.3.4 Drug Design

Drug design has been the most active area for the development of systematic procedures to identify new chemical products. Quantitative Structure Activity Relationships, QSARs, are statistically derived relationships between biological potency and physical-chemical properties. The independent physical-chemical properties represent how the various environments and reactions in the body affect the thermodynamics, transport and kinetics of drug activity.

Early QSAR applications in the pharmaceutical field were mainly in research on narcotic effects, anesthesiology, tranquilizers, sedatives, and pain-killers. Currently, QSAR methods are standard tools used by the industry in almost every drug project from initial proposals to series design to final design and testing stages[92].

Three physical-chemical properties typically used in a QSAR model are:

1. **Hammett's constant:**  $\sigma$ . Hammett proposed a constant to assign numerical values for the electronic effects of substitution on an aromatic ring[49,92]. With benzoic acid as the reference compound, this electronic parameter is defined by

the equation:

$$\sigma = \log(K_X/K_H) \quad (2.10)$$

where  $K_X$  and  $K_H$  are the ionization constants for the X-substituted and unsubstituted benzoic acid, respectively.

**2. Taft's Steric Parameter:  $E_S$ .** Taft[122] extended Hammett's idea to aliphatics by introducing a steric parameter,  $E_S$ , defined by the equation:

$$\delta E_S = \log(K_X/K_o) \quad (2.11)$$

where  $K_X$  and  $K_o$  are hydrolysis rate constants for the X-substituted derivative and unsubstituted parent compound respectively.  $\delta$  is chosen according to the system being studied. For the acid hydrolysis of esters,  $\delta$  is fixed at 1.00, and methyl was chosen as the reference system (i.e.,  $E_S$  for  $-\text{CH}_3$  equals 0.000).

**3. Partition Coefficient Ratio:  $\pi$ .** Hansch, et.al.[50] combined the concepts of Hammett and Taft to derive a hydrophobic parameter for substituents that were related more closely to biological activities. Known as  $\log P$ , this is the most popular and commonly encountered descriptor in QSAR studies. They postulated that, before it could take part in a reaction, a drug had to bind to certain target locations in the living material. To account for this drug-ligand binding capability, they proposed a hydrophobic parameter  $\pi$ , defined as:

$$\pi_X = \log P_X - \log P_H \quad (2.12)$$

where  $P_X$  and  $P_H$  are the oil-water partitioning coefficients of the X-substituted compound and of the unsubstituted parent compound, respectively. An octanol-

water system often is used to represent the fatty and the aqueous phases. In correlation equations,  $\log P_H$  is sometimes used as a constant term, and  $\log P$  is the independent variable rather than  $\pi$ .

Other descriptors derived from the electronic configuration of the molecule began to appear in QSAR studies in the late 1970's[92]. Molar refractivity(MR), which is directly connected with the electronic configuration of the molecule, has been successfully used in developing high-quality QSAR models. The use of MR in biological activity correlations originated from the suggestion that polarizability, as measured by MR, is an important aspect in the drug-ligand interaction. The strength of binding between the two was expected to relate to the resulting biological activity.

Topological indices have also played a part as independent variables. Kier and Hall's molecular connectivity indices[65],  $\chi$ , has often been used. One interesting feature of topological parameters is that they provide a quantitative way to measure the "branchedness" of a molecule.

I use one of the early studies of penicillin[51] to illustrate the QSAR approach to drug design. The initial penicillin for which derivatives were analyzed is shown in Figure 2.4. Substituents for the benzene ring were being searched for which would increase the pharmaceutical's activity.

Hansch et.al.[51] synthesized and tested a set of analogs. Table 2.17 shows these experimental results. The  $\sum \sigma$  and  $\sum \pi$  values are obtained from standard tabulations such as Hansch and Leo[53].

With this data gathered, the next step in the design was to identify a model relating

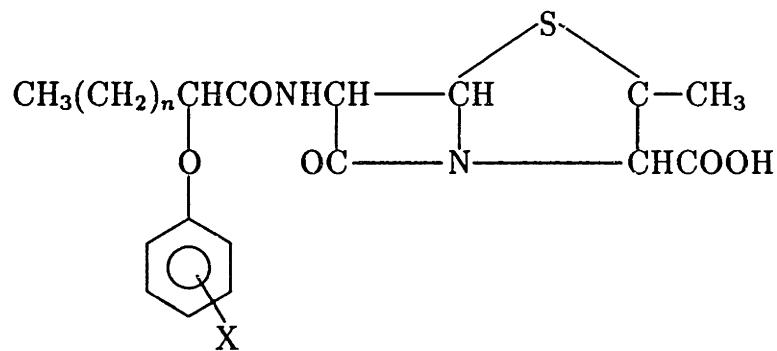


Figure 2.4: Initial Penicillin for Modification by Drug Design

Table 2.17: Input Data for QSAR Analysis

	Substituent	$\Sigma \sigma$	$\Sigma \pi$	Experimental $\log 1/C$
1	H	0.00	0.00	5.86
2	4-Cl	0.23	0.74	5.79
3	4-OCH <sub>3</sub>	-0.27	-0.04	5.69
4	$\alpha$ -Et	0.00	0.50	5.54
5	4-NO <sub>2</sub>	0.78	0.06	5.53
6	2-Cl	0.23	0.59	5.40
7	3-CF <sub>3</sub>	0.42	1.09	5.38
8	2,5-Cl <sub>2</sub>	0.60	1.35	5.24
9	$\alpha$ -Pr	0.00	1.00	5.03
10	3,5-(CH <sub>3</sub> ) <sub>2</sub>	-0.14	1.02	5.03
11	3-CF <sub>3</sub> , 4-NO <sub>2</sub>	1.20	1.15	5.03
12	$\alpha$ -Bu	0.00	1.50	5.01
13	2,4-Cl <sub>2</sub>	0.46	1.33	4.97
14	2,4-Br <sub>2</sub>	0.46	1.77	4.87
15	2,3,6-Cl <sub>2</sub>	0.83	1.94	4.72
16	4-Cyclohexyl	-0.15	2.52	4.70
17	4- <i>t</i> -Bu	-0.20	1.71	4.67
18	3,4,5-(CH <sub>3</sub> ) <sub>3</sub>	-0.31	1.54	4.65
19	4- <i>t</i> -Amyl, $\alpha$ -Et	-0.20	2.71	4.57
20	Cl <sub>5</sub>	1.43	3.44	4.25

Table 2.18: QSAR Relationships Regressed from Input Data

---


$$\log 1/C = 0.053\pi^2 - 0.610\pi + 0.019\sigma + 5.751 \quad (2.13)$$

$$r = 0.918 \quad s = 0.192$$

$$\log 1/C = 0.055\pi^2 - 0.613\pi + 5.756 \quad (2.14)$$

$$r = 0.918 \quad s = 0.187$$

$$\log 1/C = -0.445\pi + 5.673 \quad (2.15)$$

$$r = 0.909 \quad s = 0.191$$


---

activity to the physical-chemical parameters. The equations shown in Table 2.18 were obtained from a least squares fit of the data shown in Table 2.17. The correlation coefficient is represented by  $r$ .  $s$  represents the standard deviation. Table 2.19 compares the estimates made using Equation 2.15 against the experimentally derived data. The estimates show an average absolute error of 0.104 and an average absolute percent error of 2.84%.

With a quantitative relationship between structure and activity we can begin to exploit it. Comparison of Equation 2.13 with Equation 2.14 in which the  $\sigma$ -term has been dropped shows the great advantage in the use of the substituent constants  $\pi$  and  $\sigma$  to separate the electronic effect of substituents. In this instance it is quite clear that the electronic effects of the groups attached to the phenoxy ring are not important except in so far as they affect the partition coefficient of the molecule in question.

A very interesting aspect of Equation 2.15 is the negative coefficient associated with  $\pi$ . This would indicate more active derivatives could be obtained by using substituents which have negative  $\pi$  values. In this aspect the model is used as a guide to select the most appropriate set of substituents to try next.

Table 2.19: Estimated Activity for QSAR Analysis

	Substituent	Experiment	Estimate	Error <sup>†</sup>
1	H	5.86	5.67	-0.19
2	4-Cl	5.79	5.34	-0.45
3	4-OCH <sub>3</sub>	5.69	5.69	0.00
4	$\alpha$ -Et	5.54	5.45	-0.09
5	4-NO <sub>2</sub>	5.53	5.65	-0.12
6	2-Cl	5.40	5.41	-0.01
7	3-CF <sub>3</sub>	5.38	5.19	-0.19
8	2,5-Cl <sub>2</sub>	5.24	5.07	-0.17
9	$\alpha$ -Pr	5.03	5.23	-0.20
10	3,5-(CH <sub>3</sub> ) <sub>2</sub>	5.03	5.22	-0.19
11	3-CF <sub>3</sub> , 4-NO <sub>2</sub>	5.03	5.16	-0.13
12	$\alpha$ -Bu	5.01	5.01	0.00
13	2,4-Cl <sub>2</sub>	4.97	5.08	-0.11
14	2,4-Br <sub>2</sub>	4.87	4.88	-0.01
15	2,3,6-Cl <sub>2</sub>	4.72	4.81	-0.09
16	4-Cyclohexyl	4.70	4.55	-0.15
17	4- <i>t</i> -Bu	4.67	4.91	-0.24
18	3,4,5-(CH <sub>3</sub> ) <sub>3</sub>	4.65	4.99	-0.34
19	4- <i>t</i> -Amyl, $\alpha$ -Et	4.57	4.47	-0.10
20	Cl <sub>5</sub>	4.25	4.14	-0.11

<sup>†</sup> Error = Estimate - Experiment.

## 2.4 Generate and Test

Generate and test is a search paradigm using two modules: 1) generator; 2) tester[133].

The generator enumerates candidate solutions. The tester evaluates each candidate either accepting or rejecting it. When the number of candidates becomes large, exhaustive enumeration becomes impractical. Three ways to manage the search space are[54]:

1. Move the tester into the generator.
2. Prune partial solutions.
3. Abstract the search space.

Moving the tester into the generator reduces the number of incorrect candidates.

The major difficulty with this approach is that evaluating a candidate may require considerable effort. This is the case with many design problems, designing chemical processes and molecules being two examples. In both cases, a set of complex relationships are used for evaluation.

Pruning partial solutions has the difficulty in design problems that a partial infeasibility can be made feasible by the next addition to the solution. A partial molecule consisting of:



could be infeasible from structural and physical property considerations. However, adding the group

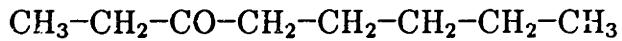


could make it feasible with respect to all tests.

Abstraction of the search space can provide a powerful means to manage the combinatorics of design problems.

## 2.5 DENDRAL

Heuristic DENDRAL[76] is a computer program which infers the molecular structure of a compound given its mass spectra. Figure 2.5 shows a mass spectrogram which, along with the empirical formula:  $C_8H_{18}O$ , are the input to DENDRAL. DENDRAL takes this input and identifies the compound being analyzed to be:



The program consists of five sections:

1. Preliminary Inference Maker
2. Data Adjustor
3. Structure Generator
4. Predictor
5. Evaluation Function

Figure 2.6 shows the relationship between these sections.

The preliminary inference maker examines a spectrum and determines what general classes of chemical substructures are confirmed or disconfirmed by the data. All hypothesized structures generated later by DENDRAL must contain all the confirmed substructures (all of which are put on a list called Goodlist); and no structure may

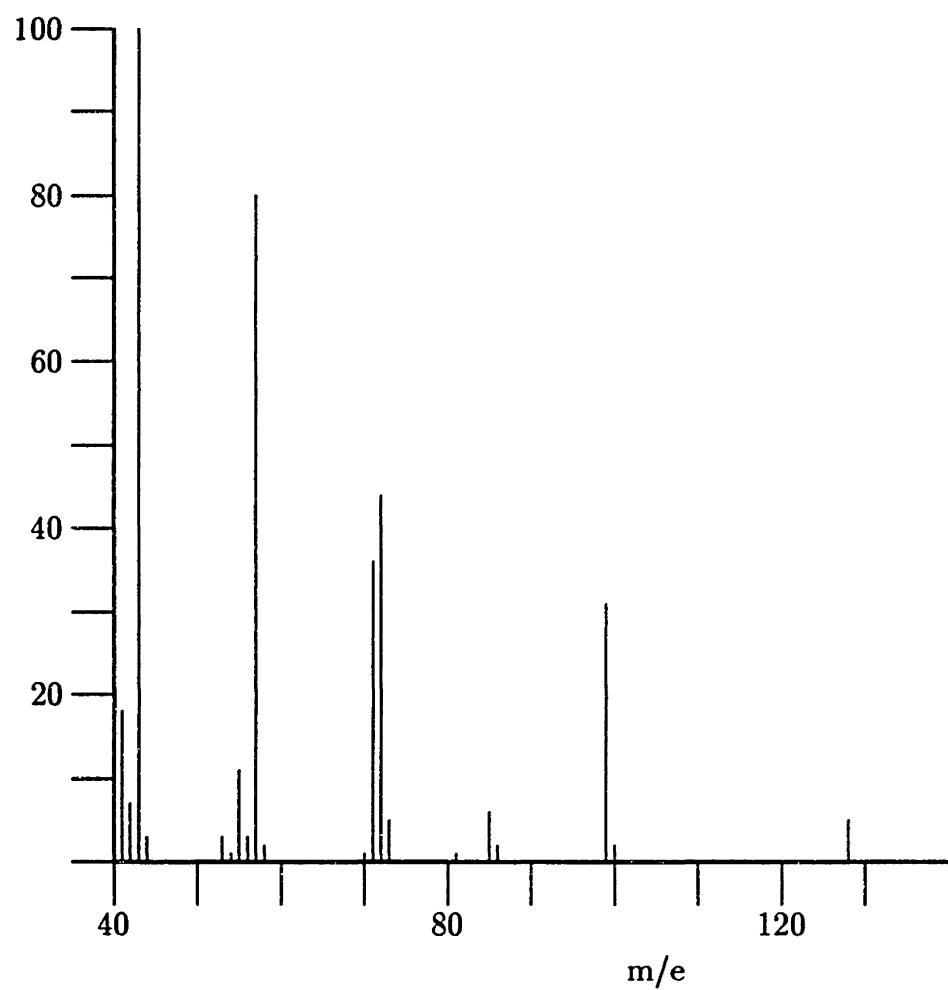


Figure 2.5: Input Data for Heuristic DENDRAL

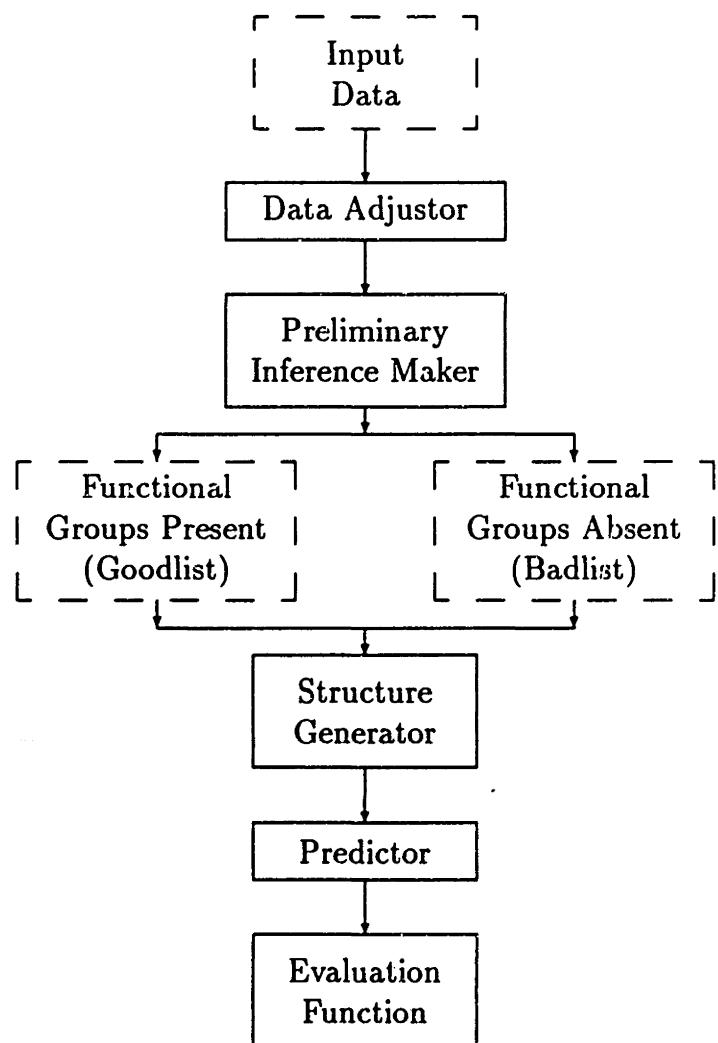


Figure 2.6: General Design of Heuristic DENDRAL's Five Major Sections

contain any substructure which is disconfirmed by the spectrum (all forbidden substructures are put on a list called Badlist.) Determination is made based upon a set of “rules”. Table 2.20 displays some of the rules used.

The *Data Adjustor* section determines which mass points of a real spectrum are significant enough to be used by later programs.

The *Structure Generator* section implements the DENDRAL algorithm[76] but with the inclusion of heuristic constraints to prevent the program from generating structures which are incompatible with chemical theory or mass spectral data. The Structure Generator takes as input:

1. a list of defined atoms with their valences and weights
2. the empirical formula
3. the mass spectrogram
4. a list of likely substructures (the goodlist)
5. a list of impossible substructures (the badlist)

The Structure Generator then generates all structures compatible with the given data.

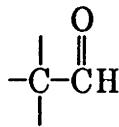
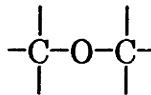
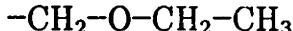
If there are no data-oriented lists of likely or impossible substructures and no spectral data, the program generates all structural variants (isomers) of the given empirical formula.

One of the important features of DENDRAL is its use of various levels of models for pruning. The *Zero-Order Theory* assumes that every bond of a structure to which it applies will break (one bond at a time) and that at least one of each pair of substructures obtained from a single break will contribute its mass to the spectrum. This Zero-Order Theory of mass spectrometry is crude but easily implemented. A more elaborate spectral theory is contained in the Predictor, but obtaining such a spectrum for an

Table 2.20: Heuristic DENDRAL's Identification Rules

Functional Group and Characteristic Subgraph	Identifying Conditions
A. Ketone	$\begin{array}{c} \text{O} \\ \parallel \\ -\text{C}- \end{array}$ <ol style="list-style-type: none"> <li>1. There are 2 peaks at mass units <math>x_1</math> and <math>x_2</math> such that:           <ol style="list-style-type: none"> <li>(a) <math>x_1 + x_2 = M + 28</math></li> <li>(b) <math>x_1 - 28</math> is a high peak</li> <li>(c) <math>x_2 - 28</math> is a high peak</li> <li>(d) At least one of <math>x_1</math> or <math>x_2</math> is high</li> </ol> </li> </ol>
B. Methyl-Ketone 3	$\begin{array}{ccccc} \text{O} & & \text{H} & & \\ \parallel & &   & & \\ \text{CH}_3-\text{C}-\text{CH}_2-\text{C} & - & \text{C}-\text{C}- & & \\   & &   & & \\ & & \text{H} & & \end{array}$ <ol style="list-style-type: none"> <li>1. Ketone conditions are satisfied</li> <li>2. 43 is a high peak</li> <li>3. 58 is a high peak</li> <li>4. <math>M - 43</math> is a low peak</li> <li>5. <math>M - 15</math> is low or possibly zero</li> </ol>
C. Ethyl-Ketone 3	$\begin{array}{ccccc} \text{O} & & \text{H} & & \\ \parallel & &   & & \\ \text{CH}_3-\text{CH}_2-\text{C}-\text{CH}_2-\text{C} & - & \text{C}-\text{C}- & & \\   & &   & & \\ & & \text{H} & & \end{array}$ <ol style="list-style-type: none"> <li>1. Ketone conditions are satisfied</li> <li>2. 57 is a high peak</li> <li>3. 72 is a high peak</li> <li>4. <math>M - 29</math> is a high peak</li> <li>5. <math>M - 57</math> is a high peak</li> </ol>
D. n-Propyl-Ketone 3	$\begin{array}{ccccc} \text{O} & & \text{H} & & \\ \parallel & &   & & \\ \text{CH}_3-\text{CH}_2-\text{CH}_2-\text{C}-\text{CH}_2-\text{C} & - & \text{C}-\text{C}- & & \\   & &   & & \\ & & \text{H} & & \end{array}$ <ol style="list-style-type: none"> <li>1. 71 is a high peak</li> <li>2. 43 is a high peak</li> <li>3. 86 is a high peak</li> <li>4. 58 appears with any intensity</li> </ol>
E. iso-Propyl-Ketone 3	$\begin{array}{ccccc} \text{CH}_3 & \text{O} & & \text{H} & \\   & \parallel & &   & \\ \text{CH}_3-\text{CH}-\text{C}-\text{CH}_2-\text{C} & - & \text{C}-\text{C}- & & \\   & &   & & \\ & & \text{H} & & \end{array}$ <ol style="list-style-type: none"> <li>1. 71 is a high peak</li> <li>2. 43 is a high peak</li> <li>3. 86 is a high peak</li> <li>4. There is no peak at 58</li> </ol>

Table 2.20 Continued: Heuristic DENDRAL's Identification Rules

Functional Group and Characteristic Subgraph	Identifying Conditions
F. Aldehyde	<ol style="list-style-type: none"> <li><math>M - 44</math> is a high peak</li> <li>44 is a high peak</li> </ol>
	
G. Ether	<ol style="list-style-type: none"> <li><math>M - 17</math> is absent</li> <li><math>M - 18</math> is absent</li> </ol>
	
H. Ether 2	<ol style="list-style-type: none"> <li>Ether conditions are satisfied</li> <li>There are two peaks at <math>x_1</math> and <math>x_2</math> such that:             <ol style="list-style-type: none"> <li><math>x_1 + x_2 = M + 44</math></li> <li>At least one of <math>x_1</math> or <math>x_2</math> is high</li> </ol> </li> </ol>
	
I. Methyl-Ether 2	<ol style="list-style-type: none"> <li>Ether 2 conditions are satisfied</li> <li>45 is a high peak</li> <li><math>M - 15</math> is low or possibly zero</li> <li><math>M - 1</math> appears with any intensity</li> </ol>
	
J. Ethyl-Ether 2	<ol style="list-style-type: none"> <li>Ether 2 conditions are satisfied</li> <li>59 is a high peak</li> <li><math>M - 15</math> appears with any intensity</li> </ol>
	
K. Primary-Amine 2	<ol style="list-style-type: none"> <li>30 is a high peak</li> <li>No other peak is high</li> </ol>
	

arbitrary structure consumes more computer time than would be practical in a program such as the Structure Generator.

The *Predictor* section contains a rough theory of mass spectrometry capable of predicting major features of mass spectra of acyclic organic molecules. The Predictor estimates a mass spectrogram for each candidate molecule.

The *Evaluation Function* now checks the degree to which the predictions agree with the original data. Evaluation is a two-step process:

1. Candidates whose predictions are inconsistent with the original data are rejected.
2. Consistent candidates are rank ordered.

## 2.6 Interval Arithmetic

The generalization of ordinary arithmetic to closed intervals is known as interval arithmetic. An interval is defined as a closed bounded set of real numbers[89]:

$$X = [\underline{X} \quad \overline{X}] = \{x | \underline{X} \leq x \leq \overline{X}\} \quad (2.16)$$

Interval analysis has found applicability in problems in which the initial data contained uncertainty or in which a range of answers are sought.

Intervals have been widely used in the field of artificial intelligence. Order of magnitude reasoning and inequality reasoning are two areas which have used intervals. In order of magnitude reasoning[9] real numbers are separated into three intervals:

$$[-\infty \quad -\epsilon] \quad [0 \quad 0] \quad [\epsilon \quad \infty]$$

These intervals are typically represented as  $-$ ,  $0$ , and  $+$ . Inequality reasoning algorithms use intervals to identify regions of constant behavior[109].

Thieler[123] presents a set of examples in which interval analysis could be used in technical calculations. Among his examples were calculations involving electric circuits, lenses, and density determination.

Himmelblau[58] applied interval analysis to the chemical engineering problem of mass balance rectification. He found that the advantage of interval analysis over the existing methods of mathematical programming, Kalman filtering, sensitivity analysis, and estimation methods were that

1. Interval analysis can accommodate problems in which more than one local solution exists
2. Interval analysis yields results in terms of intervals that are more easily understood by plant personnel than the typical confidence limits obtained from statistical analysis.

Much research in interval analysis has focused on developing procedures for reducing excess width. Moore[89] proposed an algorithm combining derivative inspection and united extension. This approach was extended by Asaithambi, et.al.[4]. Ratschek and Rokne[98] examined numerous applications of the centered form. Rall[95] used information about the monotonicity of a function using the process of automatic differentiation.

# Chapter 3

## Modeling Molecular Design

The goal of my research was to develop a systematic molecular design procedure. I codified many of the concepts discussed in Chapter 2 into a model of the molecular design process. This model is captured in a six step methodology. In this chapter I discuss each of these steps and how they were suggested by the previous work.

### 3.1 Constraint Elucidation

The molecular selection and molecular design studies discussed in Chapter 2 all begin with an elucidation of the constraints candidate molecules must satisfy. The quality of these constraints ranged from being well characterized such as in Tortorello and Kinsella's[124,125] study of polymeric coatings, to poorly characterized such as in the solvent selection constraints of finding two immiscible compounds. The first step of a systematic design is to identify those constraints which are important.

Besides physical property constraints, structural and chemical constraints must also

be considered. Gani and Brignole[13,42] specified acceptable structural characteristics of groups restricting their generation of candidate solvents. They also eliminated groups which could lead to chemical instability or corrosion problems.

### 3.2 Property Estimation

Because new molecules are being designed, experimental values for their physical properties are unknown. It is necessary to estimate their physical properties. The second step in molecular design is to identify estimation techniques for each of the physical properties occurring in our constraints. The solvent selection and design studies of Francis[38], Berg[8], Gani and Brignole[13,42], and Lo et.al.[77] show that there are often several techniques for estimating important properties.

The physical property estimation techniques I chose to concentrate on are group contribution and equation oriented estimation techniques. Group contribution estimation techniques have the ability to estimate physical properties given only a compound's molecular structure. Equation oriented estimation techniques extend the number of physical properties which can be designed for.

### 3.3 Molecule Generation

Derringer[26], Gani and Brignole[13,42], and Tortorello and Kinsella[124,125] all used a generate and test paradigm to search for new molecules. All three approaches used group contribution techniques as the basis of design. A set of groups were chosen, their important physical properties estimated, and the imposed physical property constraints

checked.

Gani and Brignole discussed the need for structural constraints to ensure the chosen groups could actually be connected together to form a feasible molecule. Derringer[26] used only groups with two free single bonds. The work involved designing the repeat unit for polymers. No structural constraints were needed since any number of groups chosen would be structurally feasible.

Gani and Brignole also discussed the need to manage the combinatorics of the group selection problem.

### **3.4 Molecule Enumeration**

The generate and test paradigm used in molecular design studies produces a collection of groups which satisfy physical property constraints. These groups can often be connected together in several ways producing a variety of molecules. This enumeration problem is analogous to that addressed by DENDRAL[76].

Enumeration of complete molecules is necessary for their final evaluation. Many accurate estimation techniques require a more global knowledge of a molecule's structure than provided by groups. Chemical constraints often require knowledge of the actual connection between groups.

### **3.5 Detailed Evaluation**

Douglas's[29] procedure for designing chemical plants begins using simplified models. This is done partly to speed pruning and partly because the information required in

more rigorous models is not known at the beginning of a design. Using rigorous models for design requires us to design a complete chemical plant before we could evaluate it. Using simplified models allows us to prune infeasible designs quickly. Only promising designs need to be rigorously evaluated.

Group contribution models are well suited for use in a generate and test procedure. However, more rigorous models like molecular modeling techniques produce more accurate estimates. At the detailed evaluation step we know the entire molecular structure of our candidate molecules. We now use rigorous estimation techniques to further check satisfaction of our physical property constraints.

Chemical constraints are also applicable at this step. Identifying undesired substructures can only be done after our collection of groups have been connected together.

The classic final evaluation step is to perform laboratory experiments on the candidate molecules. This is the most accurate evaluation method but also the most expensive. However, the design steps ensure that only the most promising candidates make it to this step.

## 3.6 My Methodology

The concepts discussed in the preceding sections were collected into a systematic molecular design methodology. My methodology consists of six steps:

1. Problem Formulation
2. Target Transformation
3. Group Selection
4. Molecule Enumerate

## 5. Molecule Screening

## 6. Final Evaluation

Each of these steps is briefly described here. Chapters 4 through 10 discuss each of these steps in detail.

**Step 1: Problem Formulation:** The first step in any design is to identify the target[118]. In molecular design our target consists of constraints on important physical properties. Molecular design targets are often stated in abstract terms such as: find a stronger polymer than kevlar; develop a freon replacement; find a solvent to facilitate the separation of acetic acid from water. Taking an abstract target and developing constraints on tensile strength, vapor pressure, and selectivity is done in this first step.

**Step 2: Target Transformation:** Often the target properties identified in the problem formulation step are not directly related to molecular structure. The purpose of the target transformation step is to propagate the target constraints to constraints on properties which are directly related to molecular structure. In my thesis I used equation oriented estimation techniques to propagate physical property constraints onto constraints involving physical properties estimated by group contribution estimation techniques. In this manner every physical property constraint is a constraint on molecular structure.

**Step 3: Generate and Test:** Two design procedures were developed based upon the generate and test paradigm. Evaluation of physical property constraints, structural constraints, and management of combinatorics were the major issues considered in the

development of the procedures. The interactive design method represents the design problem graphically enabling the designer to guide the search. The automatic design procedure uses an abstraction algorithm which efficiently searches a large number of molecules.

**Step 4: Molecule Enumeration:** The interactive and automatic design procedures produce a collection of groups which satisfy structural and physical property constraints. These groups can often be connected in several ways. All molecules which can be formed from these groups are now enumerated.

**Step 5: Molecule Screening:** Candidate molecules must be stable and adhere to chemical constraints. Candidate molecules are searched for substructures suggesting chemical instability. Those candidates found possessing one or more of these disallowed substructures are pruned.

**Step 6: Final Evaluation:** More rigorous estimation techniques are used at this step to verify that the candidate molecules satisfy our imposed physical property constraints.

# Chapter 4

## Problem Formulation

The first step in any design is to identify the target[118]. In molecular design our target consists of a set of constraints on important physical properties. Identifying important physical properties is essential to establishing the target. Although there is no systematic procedure for identifying these important properties there are a number of sources of physical property constraints which should be examined for any molecular design.

In this chapter I discuss sources of constraints. I begin by presenting targets for three example molecular designs:

1. Solvent design for solvent extraction.
2. Refrigerant design for vapor recompression refrigeration cycles.
3. Polymer Design for use as a barrier material in food packaging.

I then show how these constraints can be categorized into different sources of constraints.

Table 4.1: Important Physical Properties in Solvent Design

---

Selectivity	Distribution Coefficient
Capacity	Density
Interfacial Tension	Chemical Reactivity and Stability
Corrosiveness	Viscosity
Vapor Pressure	Freezing Point
Flammability	Toxicity

---

## 4.1 Solvent Design

The success of a liquid-liquid extraction process is strongly dependent on the selection of the most appropriate solvent[77]. Treybal[126] has listed a number physical properties which should be considered in solvent selection. These are shown in Table 4.1. Classifying these physical properties provides some insight into the source of physical property constraints for molecular design. First I briefly describe the relevance of each of these properties to solvent selection.

### Selectivity

One of the most important properties of a good solvent **S** is its ability to extract **B** from a mixture of **A** and **B** preferentially, so that the ratio of **B** to **A** in the extract after removal of solvent is different from the ratio of these components in the solvent-free raffinate[94]. Figure 4.1 shows a single-stage extraction graphically depicted on a triangular diagram. Feed **F** is contacted with solvent **S** to produce a mixture of composition **Q**. Being inside the two phase region this mixture separates into two mixtures of compositions **E** and **R**. The solvent free compositions of these mixtures are represented by **E'** and **R'**.

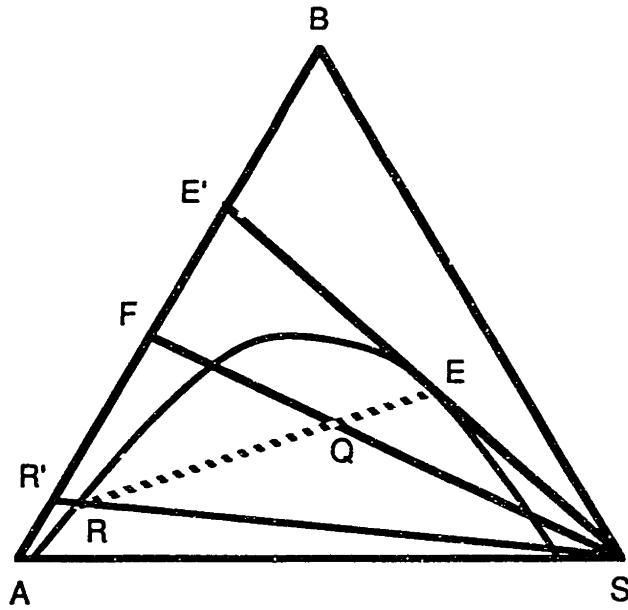


Figure 4.1: Selectivity of **S** for the Separation of **B** from **A**.

The selectivity of solvent **S** for separating **B** from **A** is given by:

$$\beta_{B,A} = \frac{x_{BS}/x_{AS}}{x_{BA}/x_{AA}} \quad (4.1)$$

where  $x_{ij}$  denotes the mole fraction of component  $i$  in the  $j$ -rich phase. At equilibrium the activities of a component in either phase must be equal. Using the definition of activity the selectivity defined in Equation 4.1 is expressed as:

$$\beta_{B,A} = \frac{\gamma_{AS}}{\gamma_{BS}} \frac{\gamma_{BA}}{\gamma_{AA}}. \quad (4.2)$$

### Distribution Coefficient

The distribution coefficient is given by:

$$m = \frac{x_{BS}}{x_{BA}} = \frac{\gamma_{BA}}{\gamma_{BS}} \quad (4.3)$$

representing the relative distribution of component **B** between the solvent-rich phase **S** and the **A**-rich phase[77].

### **Capacity**

Unless the solvent has the capacity to dissolve relatively large quantities of the preferentially extracted solute, in addition to having a high selectivity, it will likely be uneconomical to use because of the large quantity that must be circulated through the extraction system[126].

### **Density**

A difference in densities of the contacted phases is essential and should be as great as possible. Not only is the rate of disengaging of the immiscible layers thereby enhanced, but also the capacity of the contacting equipment increased[126].

### **Interfacial Tension**

The interfacial tension between immiscible phases that must be settled or disengaged should be high for rapid coalescence. Very high interfacial tension usually means that mechanical agitation is needed for carrying out the extraction[94]. This is a minor disadvantage compared with the coalescence problems which may arise from too low a value.

## **Chemical Reactivity and Stability**

Ordinarily chemical reactions between solvent and components of the feed solution, yielding products extraneous to the process, are undesirable since the yield of desired product is reduced, solvent recovery problems are increased, and losses of solvent may be incurred. On the other hand, such chemical reaction will usually increase the distribution coefficient for the reacting solute and for this reason may sometimes be sought[126].

## **Corrosiveness**

In order to reduce the cost of equipment, the solvent should cause no severe corrosion difficulties with common materials of construction or with those ordinarily used to handle the feed to the extraction process[126].

## **Viscosity**

Low viscosity results in low power requirement for pumping and agitation, rapid extraction, rapid settling of dispersions, and high heat and mass-transfer rates[126]. Solvents may sometimes be mixed with low viscosity inert diluents to improve this property.

## **Vapor Pressure**

Ordinarily low vapor pressure is desired so that storage and extraction operations are possible at atmospheric, or at most only moderately high, pressure and so that losses of solvent are kept to a minimum[126]. Exceptions to this requirement are frequently made in the interest of easy recovery of the solvent by volatilization and other desirable

properties.

### **Freezing Point**

The solvent should have a sufficiently low freezing point so that it may be conveniently stored and otherwise handled at outdoor temperatures in cold weather[126].

### **Flammability**

Low flammability is desirable for reasons of safety. The flash point is frequently used as a numerical indication of this property. If the solvent can be burned, it should have a high flash point and close concentration limits for explosive mixtures with air[126].

### **Toxicity**

Highly poisonous materials are difficult to handle industrially. Solvents that might leave toxic residues in food and pharmaceutical products must be avoided.

## **4.2 Refrigerant Design**

Refrigeration is the process of transferring heat from a low temperature to a high temperature at the expense of work. The principle methods of refrigeration available are:

1. Vapor Recompression
2. Vapor Absorption
3. Air Cycle
4. Thermo-electric.

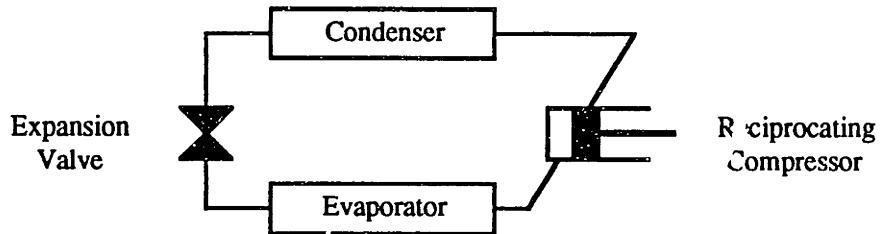


Figure 4.2: Basic Refrigeration System

The majority of plants of all sizes from domestic refrigerators to large industrial systems use the vapor recompression principle.

Figure 4.2 shows a schematic of a basic vapor recompression refrigeration system. The four major components of the cycle are a compressor, an expansion valve, and two heat exchangers. The thermodynamic processes which occur within the system are best depicted on a Pressure-Enthalpy diagram. A hypothetical P-H diagram is shown in Figure 4.3.

In the evaporator the refrigerant at low temperature is contacted with the process stream which needs to be cooled. The refrigerant is a saturated liquid at this point and is denoted by state F on Figure 4.3. As the liquid absorbs heat it evaporates. The conditions of this vapor is denoted by state A. This vapor exits the evaporator and enters the compressor. Theoretically this compression is performed adiabatically resulting a vapor at state B. This vapor is cooled isobarically to state C and condensed to a saturated liquid at state D. The heat rejected by the refrigerant is absorbed by some sink usually cooling water or air. The cooled, high pressure, saturated liquid is now flashed isenthalpically through an expansion valve. The temperature of the

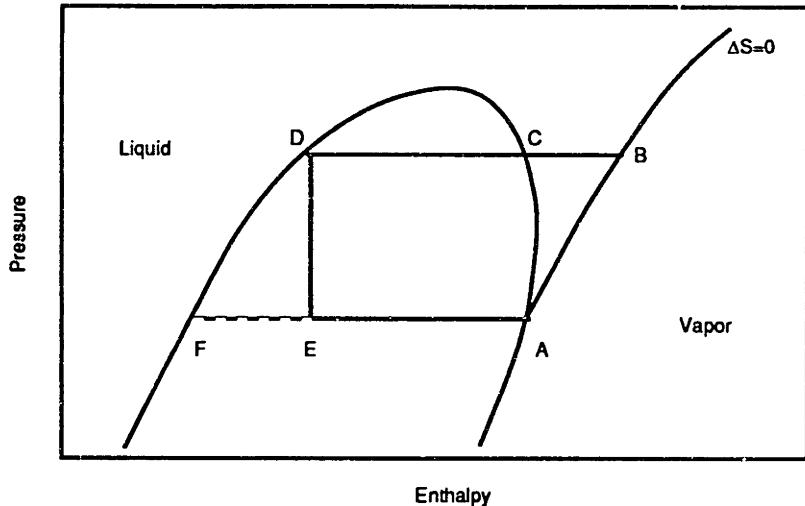


Figure 4.3: Hypothetical Refrigeration Cycle

resulting vapor-liquid mixture is lowered. The liquid at state F begins the cycle again.

Generally, the best refrigerant for a system is that which requires the smallest compressor for operation at design conditions[113]. A relationship for the volumetric flowrate of refrigerant needed for a given refrigeration load is easily obtained from thermodynamics.

If  $Q$  is the required refrigeration load and  $\Delta H_v$  is the refrigerant's molar enthalpy of vaporization at the conditions of the evaporator, then the ideal refrigerant molar flowrate is given by:

$$F_{ideal} = \frac{Q}{\Delta H_v}. \quad (4.4)$$

However, some of the refrigerant flashes upon passage through the expansion valve. This flashed vapor reduces the available refrigerant and adds to the total required refrigerant flowrate. This additional flowrate is calculated by adjusting the  $\Delta H_v$  for

this lost enthalpy:

$$\Delta H_{adj} = \Delta H_v - \Delta H_f \quad (4.5)$$

where  $\Delta H_f$  is the fraction of the latent heat expended to produce non-useful adiabatic cooling. It is a function of the heat capacity of the liquid and is given by:

$$\Delta H_f = C_{pL} (T_{condenser} - T_{evaporator}) . \quad (4.6)$$

The volumetric flowrate of the required amount of refrigerant is given by:

$$V_r = \frac{Q}{\Delta H_v - C_{pL} (T_{condenser} - T_{evaporator})} V_s \quad (4.7)$$

Thus, a good refrigerant has the characteristics of having a large enthalpy of vaporization, a small vapor volume, and a small liquid heat capacity. Additional constraints are listed in the following subsections.

### Evaporator Pressure

The pressure in the evaporator should be maintained above atmospheric to minimize the likelihood of air and moisture leaking into the cycle[28]. Douglas[29] recommends 5 psig.

### Melting Point

The melting point of the refrigerant must be well below the minimum temperature at which the system will be operated[94].

## Critical Properties

The critical temperature and critical pressure must be well above the operating temperature and pressure[94].

## Condenser Pressure

Reasonably low condensing pressures under normal atmospheric conditions are also desirable in that they allow the use of lightweight materials in the construction of the condensing equipment, thereby reducing the size, weight, and cost of equipment[28].

### 4.2.1 Graphical Constraints

Using a P-H diagram to explain a refrigeration cycle greatly enhances understanding.

Additionally the P-H diagram itself is a source of constraints. The work put into the system by the compressor is given by the difference in enthalpies between state **B** and state **A**. A steeper adiabat requires less compression work. Thus, one physical property constraint which we derive from the P-H diagram is:

$$\left(\frac{\partial P}{\partial H}\right)_S = \text{large.} \quad (4.8)$$

The liquid passing through the expansion valve flashes into a vapor and liquid phase. The vapor formed is of little refrigeration value. The amount of vapor formed is found by use of the “lever arm rule”. It is given by the relationship:

$$\text{VaporAmount} = \frac{h_E - h_F}{h_A - h_F}. \quad (4.9)$$

The steeper the phase envelop in this region the smaller the amount of vapor produced.

Thus another physical property constraint we derive from the P-H diagram is:

$$\left(\frac{\partial P}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} = \text{large.} \quad (4.10)$$

Thermodynamic relationships are used to relate these abstract physical property constraints to more common physical properties. The following derivation shows an example of reducing the constraint of Equation 4.10 to more common physical properties.

Desired characteristics:

$$\left(\frac{\partial P}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} = \text{large} \quad (4.11)$$

Using the chain rule

$$\left(\frac{\partial P}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} = \left(\frac{\partial P}{\partial T}\right)_{\substack{\text{sat} \\ \text{liq}}} \left(\frac{\partial T}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} \quad (4.12)$$

For a pure component in vapor-liquid equilibrium

$$d \ln f_V = d \ln f_L \quad (4.13)$$

expanding Equation 4.13

$$\left(\frac{\partial \ln f_V}{\partial T}\right)_P dT + \left(\frac{\partial \ln f_V}{\partial P}\right)_T dP = \left(\frac{\partial \ln f_L}{\partial T}\right)_P dT + \left(\frac{\partial \ln f_L}{\partial P}\right)_T dP \quad (4.14)$$

$$\left(\frac{\partial \ln f_L}{\partial T}\right)_P = -\frac{H^L - H^\circ}{RT^2} \quad (4.15)$$

$$\left(\frac{\partial \ln f_L}{\partial P}\right)_T = \frac{V^L}{RT} \quad (4.16)$$

After substitution and rearrangement

$$\left(\frac{dP}{dT}\right)_{\text{sat}} = \frac{H^V - H^L}{T(V^V - V^L)} \quad (4.17)$$

We know

$$\left(\frac{\partial T}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} = \frac{1}{(\partial H / \partial T)_{\substack{\text{sat} \\ \text{liq}}}} \quad (4.18)$$

$$dH = T dS + V dP \quad (4.19)$$

$$\left(\frac{dH}{dT}\right)_{\substack{\text{sat} \\ \text{liq}}} = T \left(\frac{dS}{dT}\right)_{\substack{\text{sat} \\ \text{liq}}} + V \left(\frac{dP}{dT}\right)_{\substack{\text{sat} \\ \text{liq}}} \quad (4.20)$$

$$dS = \left(\frac{\partial S}{\partial T}\right)_P dT + \left(\frac{\partial S}{\partial P}\right)_T dP \quad (4.21)$$

$$\left(\frac{dS}{dT}\right)_{\substack{\text{sat} \\ \text{liq}}} = \frac{C_{pL}}{T} + \left(\frac{\partial S}{\partial P}\right)_T \left(\frac{dP}{dT}\right)_{\substack{\text{sat} \\ \text{liq}}} \quad (4.22)$$

$$\left(\frac{\partial S}{\partial P}\right)_T = - \left(\frac{\partial V}{\partial T}\right)_P \quad (4.23)$$

Finally yielding:

$$\left(\frac{\partial P}{\partial H}\right)_{\substack{\text{sat} \\ \text{liq}}} = \frac{1}{V^L - T \left(\frac{dV}{dT}\right)_P + \frac{C_{pL} T (V^V - V^L)}{H^V - H^L}} \quad (4.24)$$

### 4.3 Barrier Polymer Design

The major use of high-barrier polymers is for food and beverage packaging. Polymeric containers have the advantages of light weight, nonshatterability (as opposed to glass), ease of disposal by incineration, and potentially lower costs[66]. The functional requirement of a package is to protect its contents from the environment over the normal shelf life of the product. Many products do not need the protection afforded by high-barrier polymers, but in the packaging of most foods and beverages protection from oxygen is of great importance as can protection from the ingress of moisture which would cause dry soluble powders to cake or a loss of moisture which may adversely affect the viscosity of water-based liquids or semiliquids. Many foods are very sensitive to oxidation which can cause flavor changes or discoloration. Tomato-based foods, such as catsup, are a prime example. For carbonated beverages both retention of carbon dioxide and protection from oxygen are important. Loss of 10% or more carbonation can be easily

detected by taste. Beer flavor is affected by oxygen levels of less than 2 ppm.

There are other important properties besides permeability and related transport properties.

### **Heat Distortion Temperature**

A low heat-distortion temperature will bar an otherwise suitable polymer from many areas of food packaging where the contents are filled at elevated temperature or processed after filling at high temperatures such as beer pasteurization[66].

### **Tensile Strength**

The bottle or package must have sufficient strength to withstand the stresses imposed during filling and processing as well as protecting its contents on its journey to the consumer. Toughness or impact resistance is equally important for the same reasons.

One of the advantages that a plastic container has over glass is its nonshatterability when accidentally dropped[66].

---

### **Creep**

Resistance to creep or cold flow is of special importance in a carbonated beverage container which is subject to high internal pressures over long periods of time. If the cold flow properties are relatively high, the bottle will distort in time causing an apparent drop in fill as well as functional problems such as rocker bottoms, etc. In addition to fill line drop and distortion, an expanding container will also cause a loss of pressure which can be far greater than that caused by permeation[66].

## Optical Clarity

Not all packaging applications require good optical clarity but this property is a decided plus especially when the competition is glass.

## 4.4 Sources of Constraints

These three example molecular design problems have very different targets. However, analyzing these targets one finds that the sources of the physical property constraints are similar. There are primarily four sources of physical property constraints:

1. Performance requirements
2. Processing restrictions
3. Equipment limitations
4. Safety and environmental concerns.

The actual classification of constraints into these four categories is not important.

These categories serve as guides for the identification of constraints.

### 4.4.1 Constraints from Equipment

Whether in processing or in final use any chemical product interacts with process equipment. The limitations of this equipment are a source for a number of physical property constraints.

In our refrigerant design example one requirement was that the compressor's adiabatic discharge temperature be below 275°F. This was to limit the possible damage which could result to the compressor seals. A second physical property constraint depends upon the compressor construction. With few exceptions, the oil required for

lubrication of the compressor is contained in the compressor's crankcase. Here the lubricant is subject to contact with the refrigerant. One important characteristic which differs for various refrigerants is oil miscibility. One of the principle effects of an oil miscible refrigerant is to dilute the oil in the crankcase. This lowers the oil's viscosity reducing its lubricating qualities[28]. Oil miscibility is not a major consideration in the selection of a refrigerant. However, since it greatly influences the design of the compressor and other system components, including the refrigerant piping, the degree of oil miscibility is an important refrigerant characteristic and should be considered in detail[28].

A second example of how equipment limitations provide physical property constraints comes from the search for working fluids for use in power cycles. The typical working fluid in a conventional Rankine cycle is steam. One of the major limitations of steam is that its critical temperature of 647.3K is well below the metallurgical limit of 894K[73]. This necessitates superheating and permits the addition of only a small amount of heat at the highest temperature of the cycle. A desired physical property of a new working fluid would thus be that it have a critical temperature above the metallurgical limit of 894K.

#### **4.4.2 Storage**

One of the most interesting physical properties of importance in solvent selection is the normal melting point. It is desirable to have  $T_m$  be above the coldest temperature experienced during the winter months. This constraint comes from considering the conditions under which solvents are stored. If the solvent's  $T_m$  is too high then heating

of storage facilities will be needed thus adding expense.

#### **4.4.3 Physical Property Constraints**

Many physical property constraints come from the economic evaluation of the process in which the compound is to be used. When optimizing a chemical operation it is common to develop an economic model which relates cost to operating conditions. Inherent in this cost model are the physical properties of the chemical compounds used in the process.

#### **4.4.4 State**

Processing conditions specify the temperatures in which the compound must be a solid or liquid. These considerations often specify constraints on  $T_b$ ,  $T_m$ ,  $T_c$ , and  $T_g$ .

### **4.5 Summary**

At the completion of the problem formulation step we have a list of constraints on the important physical properties of the chemical product. These constraints were derived from a variety of sources. Most constraints needed to be refined from constraints on abstract physical properties to constraints on well characterized physical properties.

The physical property constraints developed in the problem formulation section must now be related to molecular structure. This is done through the use of equation oriented and group contribution estimation techniques. The specification of these techniques is done in the next step of the methodology: target transformation.

# Chapter 5

## Target Transformation

The second step of my methodology prepares our target for use in the design procedures.

Two tasks must be performed at this step:

1. Choose estimation procedures for each of the physical properties present in the target constraints.
2. Find a consistent set of groups which can be used for all the chosen estimation procedures.

### 5.1 Estimation Procedures

The design procedures to be discussed in Chapters 6 and 7 require the ability to evaluate the target constraints given only the molecular structure of candidate molecules. For some physical properties, such as  $T_b$ , there are group contribution estimation techniques which directly estimate the property from a collection of groups. Other properties, such as  $P_{vp}$ , are estimated by equation oriented estimation techniques.  $P_{vp}$  is estimated by a function of other physical properties, not by groups.

I term properties which are estimated by group contribution techniques *fundamental properties*. This is to imply that no other properties are required in their estimation. Properties which are estimated by equation oriented estimation techniques are termed *non-fundamental properties*. This is to imply that other properties are required for their estimation. The actual distinction between fundamental and non-fundamental properties is dependent upon the technique chosen for estimation.  $\Delta H_{vb}$  is estimated by both group contribution techniques[62] and equation oriented estimation techniques[16, 104,130].

To relate non-fundamental properties to molecular structure it is necessary to develop an estimation procedure. This is simply a specification of estimation techniques which will allow a non-fundamental property to be estimated from group contributions.

An example estimation procedure for  $P_{vp}$  is:

- 1)  $P_{vp} = P_{vp}(T_b, T_c, P_c)$  by Riedel-Plank-Miller Technique
- 2)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  Technique
- 2)  $P_c = P_c(\text{groups})$  by Lydersen  $P_c$  Technique
- 2)  $T_c = T_c(\text{groups})$  by Fedors  $T_c$  Technique

The indentation and numbering denote the dependency of the estimation techniques.

I estimated  $P_{vp}$  using the Riedel-Plank-Miller equation oriented estimation technique[88]. This technique requires the physical properties:  $T_b$ ,  $T_c$ , and  $P_c$ . Estimation techniques must now be chosen for each of these properties.  $T_c$ ,  $T_b$  and  $P_c$  are fundamental properties.  $T_b$  is estimated by Joback's  $T_b$  group contribution estimation technique[62].  $P_c$  is estimated by Lydersen's  $P_c$  group contribution estimation technique[78].  $T_c$  is estimated by Fedors's group contribution estimation technique[35].

## 5.2 Selection Criteria

There are usually several estimation techniques available to for any physical property. Deciding which to choose is dependent upon the design method to be used. The interactive design method, Chapter 7, performs design in a physical property space which is a graph in which each of the axes corresponds to a fundamental physical property. Simply because it is not possible to display graphs of dimension greater than three, it is desirable to choose those estimation techniques which lead to a minimum number of fundamental properties. The automatic design method, Chapter 6, is not affected by the number of fundamental properties needed to be estimated. Thus, one should choose the estimation techniques of highest accuracy.

## 5.3 Restriction

In order to improve the accuracy of estimation, some estimation techniques provide corrections for certain classes of molecules. Benson's method[101] is an example of such a technique. Ring corrections are provided for certain classes of molecules.

Using these corrections requires knowledge of the complete molecular structure of our candidate molecule. We do not have this information during the design process. In the interactive design method, Chapter 7, we put the molecule together piece by piece evaluating the partial solutions. We do not know how large a ring will be formed to apply a correction. In the automatic design method, Chapter 6, we evaluate abstract molecules. We do not even know at the time of evaluation if we have ring molecules or not.

One possible way in which to use such estimation techniques is to ignore the correction terms. This provides a uniform estimation method for all molecules. If the correction terms are small then this approach should be adequate.

## 5.4 Group Consistency

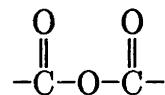
The major restriction on the choice of estimation techniques is that the chosen techniques form a consistent set of groups. The estimation procedure for  $P_{vp}$  shown previously uses three different group contribution estimation techniques. Each of these techniques has its own set of groups. The set of groups used for the entire estimation procedure is the intersection of these sets of groups. If this intersection is empty the estimation procedure is not usable.

For example, the estimation procedure shown in Table 5.1 uses two group contribution estimation techniques: gct-1 and gct-2. The example groups associated with each of these group contribution estimation techniques is shown in Table 5.2. The intersection of these group sets is:

$$( G1 \quad G4 \quad G6 )$$

Molecules can be estimated by this estimation procedure only if they are composed of these groups.

For some group sets it is possible to find inclusion sets. The group:



is contained in the group set for Franklin's group contribution estimation technique[40]

---

Table 5.1: An Example Estimation Procedure

---

- 1)  $PP_1 = PP_1(PP_2, PP_3)$  by EOT-1
  - 2)  $PP_2 = PP_2$  by GCT-1
  - 2)  $PP_3 = PP_3$  by GCT-2
- 

but it is not contained in Lydersen's set[78]. However, Lydersen's set does contain the groups:



which can be joined to form the first group. The contribution for this new Lydersen group is simply the sum of the contributions of the two included groups.

If the number of groups intersected and included is a small fraction of the original group sets then the use of the two techniques will drastically reduce the variety of molecules which can be designed. One possible solution is to regress the contributions of one group contribution technique onto the groups of another.

---

Table 5.2: Group Sets for GCT-1 and GCT-2

---

GCT-1	GCT-2
G1	G1
G2	G4
G3	G6
G4	G7
G5	G8
G6	G9

---

A linear group contribution estimation technique can be represented in the form:

$$PP = c + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_i \quad (5.1)$$

or in vector notation

$$PP = c + \underline{N} \underline{\Delta} \quad (5.2)$$

To form the contributions,  $\Delta_{new}$ , for a new group set,  $N_{new}$ , from an old group set,  $N_{old}$ , we begin by equating the physical property values estimated from Equation 5.2 for each group set:

$$c + \underline{N}_{new} \underline{\Delta}_{new} = PP = c + \underline{N}_{old} \underline{\Delta}_{old}. \quad (5.3)$$

Solving for  $\Delta_{new}$ , we obtain:

$$\underline{\Delta}_{new} = \left( \underline{N}_{new}^t \underline{N}_{new} \right)^{-1} \underline{N}_{new}^t \underline{N}_{old} \underline{\Delta}_{old} \quad (5.4)$$

To solve Equation 5.4 it is necessary to form the vectors  $\underline{N}_{new}$  and  $\underline{N}_{old}$ . This can easily be done by simply selecting a large set of molecules and dividing each into groups based on the two different group sets.

# Chapter 6

## Automatic Design

The automatic design procedure exhaustively searches a large number of molecular structures for compounds possessing properties satisfying our target. Implementing this search procedure required representing molecular structure, physical properties, and structural constraints in the computer. The computer generates candidate molecular structures, estimates physical properties, tests these properties against our target, and tests the candidate molecule for structural feasibility.

In this chapter I describe the generate and test AI paradigm which is the basis of the automatic procedure. I describe the knowledge required for the procedure and how this knowledge is represented in the computer. To overcome the combinatoric explosion associated with the design I abstract the search space. This abstraction is done using interval arithmetic. I introduce some of the concepts of interval arithmetic and show how it is used to compute ranges of physical properties. I finally summarize the automatic design procedure showing how the molecular knowledge, generate and test paradigm, search space abstraction, and interval computations are integrated into

an overall design procedure.

## 6.1 Generate and Test

The *generate and test* search paradigm uses two basic modules. One module, the *generator*, enumerates possible solutions. The second, the *tester*, evaluates each proposed solution, either accepting it or rejecting it.

Depending on the purpose and the nature of the problem, the generator may generate all possible solutions before the tester takes over, or alternatively, generation and testing may be interdigitated. Search may stop when one acceptable solution is found, or search may continue until some satisfactory number of solutions is found, or search may continue until all possible solutions are generated and tested.

Good generators have three qualities[133]:

1. Good generators are complete. They eventually produce all possible solutions.
2. Good generators are nonredundant. They never damage efficiency by proposing the solution twice.
3. Good generators are informed. They use possibility-limiting information, restricting the solutions they propose accordingly.

## 6.2 The Generator

I represent molecules as collections of groups. Chloropropane is represented as:



The generator constructs molecules by selecting a collection of groups from an initial set. Table 6.1 shows an example initial set of groups.

Table 6.1: Initial Set of Groups

>CH <sub>3</sub>	-CH <sub>2</sub> -	>CH-	>C<
=CH <sub>2</sub>	=CH-	=C<	=C=
≡CH	≡C-	-F	-Cl
-Br	-I	-OH	-O-
>CO	-CHO	-COOH	-COO-
=O	-NH <sub>2</sub>	>NH	>N-
-CN	-NO <sub>2</sub>	-SH	-S-

The combinations of groups which can be selected is infinite. However, from practical considerations, molecules for a typical application fall within some size range which can be translated into an upper limit on the number of groups chosen. For example, refrigerants are generally of a small molecular weight. Placing a limit of 10 on the number of groups which can be used to form a molecule is a conservative bound. A lower limit is established from the fact that at least two groups must be used to form a structurally feasible molecule.

With limits on the minimum and maximum number of groups which can be chosen the generator begins selecting groups. The generator begins by selecting all combinations of 2 groups from the initial set. Choosing all possible combinations of 2 groups is accomplished by generating all *compositions* of 2 over the number of groups in the initial set. A composition of a number is defined as the set of positive integers whose sum equals the number[105]. Thus the compositions of 3 are (3), (2 1), (1 2), and (1 1 1). The integers collected to form a composition are called its parts, and the number which is the sum of these parts is the composed number. I extend the composition of a number by restricting the number of parts and allowing the inclusion of zeros.

Thus the compositions of 3 restricted to 2 parts are (3 0), (2 1), (1 2), and (0 3).

Thus to generate all combinations of 2 groups from the initial set we begin with a vector of the 44 initial groups:

$$(-\text{CH}_3 \quad -\text{CH}_2- \quad -\text{CH}_2- \quad >\text{CH}- \quad \dots \quad -\text{SH} \quad -\text{S}-)$$

We then generate all compositions of 2 restricted to 44 parts:

$$(2 \ 0 \ 0 \ \dots \ 0 \ 0)$$

$$(0 \ 2 \ 0 \ \dots \ 0 \ 0)$$

⋮

$$(1 \ 1 \ 0 \ \dots \ 0 \ 0)$$

$$(1 \ 0 \ 1 \ \dots \ 0 \ 0)$$

⋮

$$(0 \ 0 \ 0 \ \dots \ 1 \ 1)$$

### 6.2.1 Molecules

Each of the restricted compositions generated above represents a molecule. For example the first restricted composition above:

$$(2 \ 0 \ 0 \ \dots \ 0 \ 0)$$

corresponds to selecting two  $-\text{CH}_3$  groups from the initial set. This composition thus represents the ethane molecule.

Table 6.2: Example Candidate Molecules

1)	$-\text{CH}_3, -\text{CH}_3, -\text{CH}_2-$	1.256
2)	$-\text{CH}_3, -\text{CH}_3, -\text{CH}_3$	1.215
3)	$-\text{CH}_3, -\text{CH}_3, -\text{O}-, -\text{O}-, -\text{O}-$	2.659
4)	$-\text{CH}_3, >\text{C}<, >\text{N}-, -\text{F}, -\text{F}, -\text{F}, -\text{F}$	3.010
5)	$\equiv\text{CH}, \equiv\text{C}-, -\text{Cl}$	1.017
6)	$>\text{C}<, -\text{F}, -\text{F}, -\text{F}, -\text{F}$	14.767

## 6.3 The Tester

A candidate molecule is generated by selecting a collection of groups from this initial set. Some possible candidates are given in Table 6.2. This selection procedure corresponds to the generator of our search paradigm. The tester consists of constraints used to prune infeasible candidate molecules. I use three types of constraints:

1. Property Constraints
2. Structural Constraints
3. Chemical Constraints

Each of these constraints uses different information obtained from the candidate molecules.

### 6.3.1 Property Constraints

The estimation procedures established during target transformation are used to test each candidate molecule for satisfaction of the target constraints. For example, I use the constraint:

$$P_{vp}(T = 273 \text{ K}) > 1.01 \text{ bar.} \quad (6.1)$$

For each of the candidate molecules in Table 6.2 I use the estimation procedure

- 1)  $P_{vp} = P_{vp}(T_b, T_{br}, P_c)$  by Riedel-Plank-Miller EOT
- 2)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(\text{groups})$  by Joback  $T_{br}$  GCT
- 2)  $P_c = P_c(\text{groups})$  by Joback  $P_c$  GCT

to determine each candidate molecule's vapor pressure at 273K. This vapor pressure value is then used to evaluate the constraint in Equation 6.1 for each of the candidates. Table 6.2 shows that all candidates possess vapor pressure values which satisfy the specified constraint.

### 6.3.2 Structural Constraints

Candidate molecules must also satisfy structural feasibility constraints. The groups of the candidate molecule must be able to be connected in some manner to form a complete molecule. Although Table 6.2's candidate molecule 2 satisfies our vapor pressure constraint there is no way the three groups can be connected to form a structurally feasible molecule.

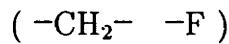
Structural constraints determine whether or not a collection of groups form a structurally feasible molecule. Three criteria define structural feasibility:

1. All groups in the collection should be able to be connected into a single connected component. The collection of groups



is not feasible by this definition because it does not form a single molecule.

2. The single connected component formed from a set of groups can not have any unconnected bonds. Connecting the groups:



gives us a single connected component with one single bond unconnected.

3. The connections made in the single connected component formed from a set of groups must all be between bonds of the same type. Single bonds may only connect with single bonds, double with double, etc.

Knowledge about restrictions on the connection of groups must be given to the system in order for it to evaluate the structural feasibility of candidate molecules. I examined a variety of approaches to systematically identifying structural constraints but was unsuccessful in finding any. Graph theory provided some of the constraints on relationships which must hold for a feasible graph. However, the central question of being able to decide whether or not a set of nodes with specified edges could form a feasible graph was not addressed.

I organized structural constraints by “Knowledge Class”. I describe each of the constraints I have identified.

### ***Known: Existence***

If the group contributions under investigation do not contain complete molecules we know that at least two groups are needed to form a feasible molecule.

**Structural Constraint 1** *If G is a collection of n groups, then n ≥ 2.*

## Known: Ring Class

Knowledge of ring class gives us some constraints on the number and types of groups which must be present either to form or not form rings. The ring class of a group can be one of three possibilities:

1. Acyclic
2. Cyclic
3. Mixed.

This distinction was made by Gani and Brignole[42]. Each bond of a group is denoted as to whether or not it can be included in a ring. Groups which have all their bonds required to be in a ring are denoted cyclic groups. Groups which have all their bonds required not to be in a ring are denoted acyclic groups. Groups which have some bonds which must be in rings and some which can not are denoted as mixed groups.

The first constraint expresses the requirement that if a candidate molecule has some acyclic groups and some cyclic groups then needs some mixed groups to be feasible. Without the mixed groups there is no way the cyclic and acyclic groups can connect together.

**Structural Constraint 2** *If  $G$  is a collection of  $n$  groups with  $n_c$  cyclic groups,  $n_m$  mixed groups, and  $n_a$  acyclic groups, and*

$$n_a > 0 \text{ and } n_c > 0 \quad (6.2)$$

*then*

$$n_m > 0. \quad (6.3)$$

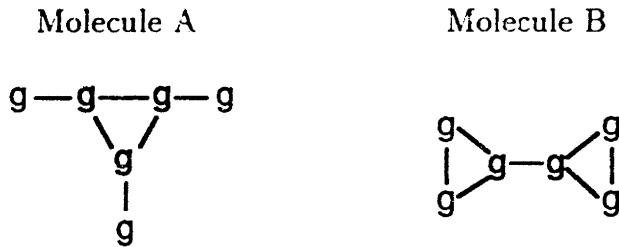


Figure 6.1: Molecules Demonstrating Structural Constraint 3

Mixed groups can not solely be used to form a feasible molecule. Either cyclic or acyclic groups are needed in addition to mixed groups. Figure 6.1 shows two molecules which satisfy the following constraint.

**Structural Constraint 3** *If  $G$  is a collection of  $n$  groups with  $n_c$  cyclic groups,  $n_m$  mixed groups, and  $n_a$  acyclic groups, then*

$$n_m > 0 \quad (6.4)$$

*implies that*

$$n_a > 0 \quad \text{or} \quad n_c > 0. \quad (6.5)$$

Molecule *A* of Figure 6.1 contains only mixed and acyclic groups while molecule *B* contains only mixed and cyclic groups.

Since the presence of either mixed groups or cyclic groups implies that at least one ring will be formed we can constrain the minimum number of ring forming groups which must be present. To form a ring requires at least three atoms. Groups which have single atoms participating in the ring require an occurrence of at least three. Even if groups can have multiple atoms contributing to the ring structure, a minimum of two groups is required.

**Structural Constraint 4** *If  $G$  is a collection of  $n$  groups with  $n_c$  cyclic groups and  $n_m$  mixed groups then either*

$$n_c + n_m \geq 3 \quad (6.6)$$

*or*

$$n_c + n_m \geq 2. \quad (6.7)$$

**Known: Global Valence**

The global valence of a group is the number of bonds associated with the group regardless of the type of bonds. With only the global valence known, our groups are identical with the nodes in a mathematical graph. The bonds connecting the groups are analogous to the edges of a graph. The analogy between molecules and graphs has long been exploited[5]. With the restrictions I have placed on how groups can connect, the molecules formed from my groups are analogous to simple graphs. This analogy lets me draw the next three structural constraints from graph theoretic constraints on simple graphs.

**Structural Constraint 5** *If  $G$  is a collection of groups, then the number of groups having an odd number of free bonds must be even.*

**Structural Constraint 6** *If  $G$  is a collection of  $n$  groups with  $b$  free bonds, then*

$$\frac{b}{2} \geq n - 1 \quad (6.8)$$

**Structural Constraint 7** *If  $G$  is a collection of  $n$  groups with  $b$  free bonds, then*

$$\frac{b}{2} \leq \frac{1}{2}n(n - 1). \quad (6.9)$$

### *Known: Bond Type*

There are five bond types:

Single      Double      Triple      Ring-Single      Ring-Double

The bond type of a group does not identify the occurrence of each type. Thus the group:

$=C<$

has the bond types *double* and *single*. This information leads us to the following constraint.

**Structural Constraint 8** *If a collection of groups contains more than one bond type then there must be a transition group containing each bond type. A transition group is one which contains more than one bond type.*

For example if our collection of groups contains *single*, *double*, and *triple* bond types then there must be either one group containing all three bond types or two groups containing two different subsets of bond types.

### *Known: Global Valence and Ring Class*

Even though the number of bonds of a mixed group which are in a ring are unknown at this level of knowledge, we know that a minimum of 2 bonds must be involved. The remaining bonds can either attach to acyclic groups or other mixed groups. If we assume that all the remaining bonds must connect to acyclic groups then we can derive an upper limit on the number of acyclic terminators which must be present for a structurally feasible molecule. This limit is shown in the following constraint.

**Structural Constraint 9** *If  $G$  is a collection of groups with  $n_{a,i}$  denoting the number of acyclic groups with a valence  $i$  and  $v_{m,j}$  denoting the valence of some  $j^{th}$  mixed group, then*

$$n_1 \leq \sum_{\text{mixed}} (v_{m,j} - 2) + n_{a,3} + 2n_{a,4} + \dots + (i - 2)n_{a,i} + \dots \quad (6.10)$$

If the number of mixed groups and cyclic groups is zero then we have the potential of forming only acyclic molecules. Acyclic molecules are analogous to trees in graph theory. Using this analogy I derive the following constraint.

**Structural Constraint 10** *If  $G$  is a collection of  $n$  groups with  $n_i$  denoting the number of groups with a global valence  $i$  and all  $n$  groups are acyclic then*

$$n_1 = 2 + n_3 + 2n_4 + \dots + (i - 2)n_i + \dots \quad (6.11)$$

### Bond Valence

The bond valence of a group contains the information about the occurrence of each bond type. Thus the group:

=C<

has the bond valence of *single 2* and *double 1*. This information gives us the following constraint.

**Structural Constraint 11** *The number of occurrences of each bond type in a collection of groups must be even.*

If the compound contains more than 2 consecutive etherial oxygens

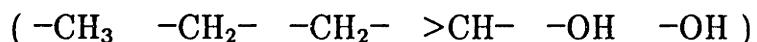
Then the compound will decompose

Figure 6.2: Example Chemical Constraint

### 6.3.3 Chemical Constraints

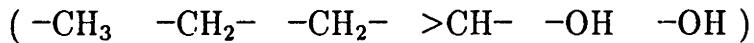
The third type of constraint is chemical stability. The molecule must retain its molecular structure when used for its desired purpose. Table 6.2's molecule 3 satisfies both physical property and structural constraints but is chemically unstable. DENDRAL used chemical stability constraints to prune its space of candidate molecules. These constraints were posed as disallowed molecular substructures. Table 9.1 lists these substructures. Figure 6.2 shows how chemical constraints can be posed in the form of a production rule.

Using these chemical constraints requires knowledge of the complete molecular structure of the candidate molecule. Given a set of groups it is first necessary to enumerate all possible structures before chemical constraints are tested. To determine whether or not the candidate molecule:



satisfies chemical constraints we first enumerate the possible ways the groups can be connected. This results in two molecules shown in Figure 8.1. Identifying the disallowed chemical substructure in molecule 2 signals that it is not chemically stable. However, since molecule 1 satisfies our criteria for chemical stability we would say that the

candidate molecule:



does satisfy our chemical stability constraints.

Some chemical constraints can be posed so enumeration is not necessary. One such constraint could be:

$$\frac{\# \text{ etherial oxygens}}{\# \text{ groups}} < 50\% \quad (6.12)$$

I decided that the majority of chemical constraints are applicable only after enumeration. Thus, instead of using chemical constraints during the design stage I use them in the final molecule screening stage. I describe the Molecule Screening stage of the methodology in Section 9.

## 6.4 Algorithm

My basic algorithm thus begins with a consistent set of groups. Collections of groups are chosen from this set. These collections are our candidate molecules. Each candidate molecule is checked to see if it satisfies property and structural constraints. If the candidate molecule satisfies these constraints then it is proposed as a solution to our design problem. Figure 6.3 shows a schematic of the automatic design procedure.

The algorithm can thus exhaustively search a large number of molecules. The major limitation is that the search could take a very long time.

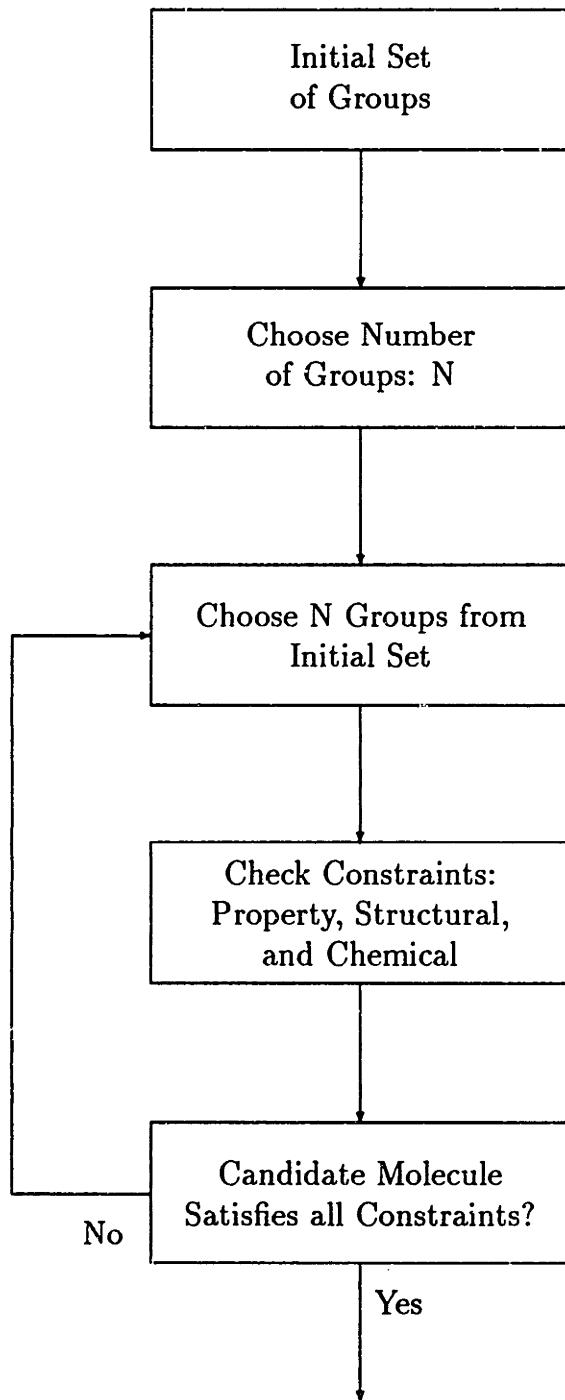


Figure 6.3: Automatic Design Algorithm Structure

Table 6.3: Combinatorics of Group Selection

k = 20 groups		k = 40 groups	
$n_{max}$	Combinations	$n_{max}$	Combinations
4	10,605	4	135,710
5	53,109	5	1,221,718
6	230,209	6	9,366,778
7	888,009	7	62,891,458
8	3,108,084	8	377,348,953
9	10,014,984	9	2,054,455,593

## 6.5 A Combinatorial Explosion

Given a reasonable number of groups in our initial set, the number of candidate molecules which can be generated is extremely large. To ensure an exhaustive search our generator begins by selecting all combinations of two groups from the initial set. When all possibilities have been exhausted the generation is repeated selecting 3-combinations.

Allowing repetition of groups and ignoring the order of selection, the number of candidate molecules which can be generated by selecting  $n$  groups from a set of  $k$  groups is given by:

$$C^R(k, n) = \frac{(k + n - 1)!}{n!(k - 1)!} \quad (6.13)$$

The total number of candidate molecules which can be selected from a set of  $k$  groups in which each candidate molecule has between 2 and  $n_{max}$  groups is given by:

$$\text{Total Candidates} = \sum_{n=2}^{n_{max}} C^R(k, n) = \sum_{n=2}^{n_{max}} \frac{(k + n - 1)!}{n!(k - 1)!} \quad (6.14)$$

Table 6.3 shows how this total number of candidates quickly grows very large.

This combinatoric problem arises from the large number of groups in the initial set. I reduce the combinatoric problem by reducing the number of groups. This is

done by abstracting the groups into clusters of groups. I call these clusters of groups *Meta-Groups*.

## 6.6 Meta-Groups

One approach to improving the efficiency of the generate and test paradigm or any search method is to abstract the search space[54]. Table 6.4 shows the groups of Table 6.1 clustered into four meta-groups.

Meta-groups can be arbitrary collections of groups or can contain groups all with a similar characteristic. When a meta-group contains groups with the same value for a characteristic, the constraints used by the tester can be applied to this abstract set of groups. In this manner we are able to reason about the entire sets of groups instead of individual groups. The combinatorics of the problem is thus reduced.

Instead of generating molecules by choosing from an initial set of groups we choose from an initial set of meta-groups. The candidate molecules formed from a collection of meta-groups are meta-molecules.

## 6.7 Meta-Molecules

Meta-Molecules are sets of molecules. The meta-molecule (2 1 0 0) using the meta-groups from Table 6.4 is the set of all molecules which can be formed by taking any two groups from Meta-Group 1 and any one group from Meta-Group 2. The number

Table 6.4: Example Meta-Groups

---

<b>Meta-Group 1</b>	$\left\{ \begin{array}{ccccccc} -\text{CH}_3 & =\text{CH}_2 & \equiv\text{CH} & -\text{F} & -\text{Cl} & -\text{Br} \\ -\text{I} & -\text{OH} & -\text{CHO} & -\text{COOH} & =\text{O} & -\text{NH}_2 \\ -\text{NO}_2 & -\text{CN} & -\text{SH} & & & \end{array} \right\}$
<b>Meta-Group 2</b>	$\left\{ \begin{array}{ccccccc} >\text{CH}_2 & =\text{CH}- & =\text{C}= & \equiv\text{C}- & >\text{CO} & -\text{COO}- \\ -\text{O}- & >\text{NH} & -\text{S}- & & & \end{array} \right\}$
<b>Meta-Group 3</b>	$\left\{ =\text{C}< \quad >\text{CH}- \quad >\text{N}- \right\}$
<b>Meta-Group 4</b>	$\left\{ >\text{C}< \right\}$

---

of molecules contained in meta-molecule (2 1 0 0) is:

$$C^R(15, 2) \times C^R(9, 1) = \frac{(15 + 2 - 1)!}{2!(15 - 1)!} \times \frac{(9 + 1 - 1)!}{1!(9 - 1)!}$$

or 1080 molecules.

As long as all the groups within each meta-group have a consistent molecular characteristic such as global valence, the structural constraints are still applicable. Meta-Groups 1 and 2 are consistent in ring class and global valence. Structural constraint 10 is thus applicable. Applying the constraint to meta-molecule (2 1 0 0):

$$2 = 2 + 0 + 2(0) = 2$$

we see that it is satisfied.

It is apparent that the manner in which we collect groups into meta-groups determines which structural constraints are applicable. The allocation done for the meta-groups in Table 6.4 does not differentiate on the basis of bond type or bond valence. We are not able to use those constraints which require this knowledge. This is the reason I presented the structural constraints organized by knowledge class.

Table 6.5:  $T_b$  Group Contributions for Meta-Group 2

Groups	Contribution
$-\text{CH}_2-$	22.88
$=\text{CH}-$	24.96
$=\text{C}=$	26.15
$\equiv\text{C}-$	27.38
$-\text{O}-$ (nonring)	22.42
$>\text{CO}$ (nonring)	76.75
$-\text{COO}-$ (ester)	81.10
$>\text{NH}$ (nonring)	50.17
$-\text{S}-$ (nonring)	68.78

## 6.8 Meta-Contributions

Associated with each of the groups in a meta-group is a contribution toward a particular physical property. The contribution the meta-group has is called a meta-contribution.

Table 6.5 shows the contributions for each of the groups in Table 6.4's Meta-Group 2 toward  $T_b$ . The meta-contribution for Meta-Group 2 toward  $T_b$  is thus the set:

$$(22.42 \quad 22.88 \quad 24.96 \quad 26.15 \quad 27.38 \quad 50.17 \quad 68.78 \quad 76.75 \quad 81.10)$$

To use meta-contributions in the calculation of physical properties it is necessary to find a representation which can be manipulated by mathematical operators. I use interval numbers.

### 6.8.1 Intervals

The generalization of ordinary arithmetic to closed intervals is known as interval arithmetic. An interval is defined as a closed bounded set of real numbers[89]:

$$X = [\underline{X} \quad \overline{X}] = \{x | \underline{X} \leq x \leq \overline{X}\} \quad (6.15)$$

Thus, intervals have a *dual* nature as both a number and a set. The basic interval arithmetic operations are:

$$\begin{aligned}
 [\underline{X} \ \ \overline{X}] + [\underline{Y} \ \ \overline{Y}] &\equiv [\underline{X} + \underline{Y} \ \ \overline{X} + \overline{Y}] \\
 [\underline{X} \ \ \overline{X}] - [\underline{Y} \ \ \overline{Y}] &\equiv [\underline{X} - \overline{Y} \ \ \overline{X} - \underline{Y}] \\
 [\underline{X} \ \ \overline{X}] * [\underline{Y} \ \ \overline{Y}] &\equiv [\min(\underline{X} * \underline{Y}, \underline{X} * \overline{Y}, \overline{X} * \underline{Y}, \overline{X} * \overline{Y}) \\
 &\quad \max(\underline{X} * \underline{Y}, \underline{X} * \overline{Y}, \overline{X} * \underline{Y}, \overline{X} * \overline{Y})] \\
 [\underline{X} \ \ \overline{X}] \div [\underline{Y} \ \ \overline{Y}] &\equiv [\underline{X} \ \ \overline{X}] * [1/\overline{Y} \ \ 1/\underline{Y}] \quad \text{iff } 0 \notin [\underline{Y} \ \ \overline{Y}]
 \end{aligned}$$

If an interval has an upper or lower bound which is zero then the reciprocal of an interval is computed accounting for the *direction* of the interval. Thus the interval  $[0 \ 10]$  is considered to contain only positive values and thus has a reciprocal of  $[0.1 \ \infty]$ . Likewise the interval  $[-10 \ 0]$  is considered to contain only negative values and thus has the reciprocal  $[-\infty \ -0.1]$ .

The meta-contribution for Meta-Group 2 in interval representation is  $[22.42 \ 81.10]$ .

## 6.9 Meta-Properties

The meta-contributions for each of the meta-groups displayed in Table 6.4 are shown in Table 6.6. The meta-contributions are listed for  $T_b$ ,  $T_{br}$ , and  $\Delta H_{vb}$ [62]. Using the group contribution estimation models[62]:

$$T_b = 198.18 + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,T_b} \quad (6.16)$$

$$T_{br} = 0.584 + 0.965 \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,T_{br}} - \left( \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,T_{br}} \right)^2 \quad (6.17)$$

Table 6.6: Meta-Contributions

Group	$T_b$	$T_{br}$	$\Delta H_{vb}$
Meta-Group 1	[-10.50 169.09]	[0.0027 0.0791]	[-0.670 19.537]
Meta-Group 2	[22.42 81.10]	[0.0020 0.0481]	[2.205 9.633]
Meta-Group 3	[11.74 24.14]	[0.0117 0.0169]	[1.691 2.138]
Meta-Group 4	[18.25 18.25]	[0.0067 0.0067]	[0.636 0.636]

$$\Delta H_{vb} = 15.30 + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,\Delta H_{vb}} \quad (6.18)$$

we can estimate the values of  $T_b$ ,  $T_{br}$ , and  $\Delta H_{vb}$  for our meta-molecule (2 1 0 0):

$$\begin{aligned} T_b &= 198.18 + 2[-10.50 \quad 169.09] + [22.42 \quad 81.10] \\ &= [199.6 \quad 617.46] \text{ K} \end{aligned} \quad (6.19)$$

$$\begin{aligned} T_{br} &= 0.584 + 0.965(2[0.0027 \quad 0.0791] + [0.0020 \quad 0.0481]) \\ &\quad - (2[0.0027 \quad 0.0791] + [0.0020 \quad 0.0481])^2 \\ &= [0.549 \quad 0.783] \end{aligned} \quad (6.20)$$

$$\begin{aligned} \Delta H_{vb} &= 15.30 + 2[-0.670 \quad 19.537] + [2.205 \quad 9.633] \\ &= [16.165 \quad 64.007] \text{ kJ/mol} \end{aligned} \quad (6.21)$$

These intervals span the range of physical property values possessed by each of the 1080 molecules in meta-molecule (2 1 0 0).

Interval values for these fundamental properties can be used in equation oriented estimation techniques. To estimate the enthalpy of vaporization at 250K for our meta-molecule we can use the Watson relation[131]:

$$\Delta H_v = \Delta H_{vb} \left( \frac{1 - T/T_c}{1 - T_{br}} \right)^{0.38} \quad (6.22)$$

$T_c$  is obtained from:

$$T_c = \frac{T_b}{T_{b_r}} = \frac{[199.6 \quad 617.46]}{[0.549 \quad 0.783]} = [254.9 \quad 1124.7]$$

Inserting into Equation 6.22 we obtain:

$$\Delta H_v = [16.165 \quad 64.007] \left( \frac{1 - 250/[254.9 \quad 1124.7]}{1 - [0.549 \quad 0.783]} \right)^{0.38} = [0.688 \quad 229.40] \text{ kJ/mol}$$

Interestingly if we used another form of the Watson relation:

$$\Delta H_v = \Delta H_{vb} \left( \frac{T_c - T}{T_c - T_b} \right)^{0.38} \quad (6.23)$$

we would divide by:

$$T_c - T_b = [254.9 \quad 1124.7] - [199.6 \quad 617.46] = [-362.56 \quad 925.1]$$

Since the interval denominator includes zero our resulting value for  $\Delta H_v$  would be  $[-\infty \quad \infty]$ . Although this interval is correct it contains a great deal of excess width.

### 6.9.1 Excess Width

The most typical complaint voiced about interval arithmetic is that its results are too conservative. Mathematically stated the interval extension of a function may produce an interval whose width is larger than the true width. The most obvious example is given by the function shown in Equation 6.24.

$$f(x) = x - x \quad (6.24)$$

Inserting a value of  $[0 \quad 1]$  for  $x$ , the value of  $f(x)$  is  $[-1 \quad 1]$ . Although this interval result includes the actual values of  $[0 \quad 0]$  it is undesirably wide.

## 6.9.2 Causes of Excess Width

The dominant cause of excess width arises from the multiple occurrences of variables in the interval extension of a function. The excess width is generated because interval mathematics treats each occurrence of these variables independently. Thus the interval result of Equation 6.24 is correct if we consider it to be:

$$f(x, y) = x - y \quad (6.25)$$

with

$$x = [0 \quad 1] \quad y = [0 \quad 1]$$

This excess width which arises from the consideration of the dependency of variables has been termed *dependency width*[106].

Other factors which contribute to the excess width are associated with the approximation of arithmetic operations on the computer due to its finite accuracy and to the approximations to functions such as log, sin, etc. For intervals of any width significantly larger than the accuracy of the machine the excess width is essentially dependency width.

The variable dependency which causes excess width can also arise from functional dependency. An example of this is the definition of the reduced boiling point:

$$T_{br} = \frac{T_b}{T_c}. \quad (6.26)$$

The typical ranges for  $T_b$  and  $T_c$  are:

$$T_b = [5 \quad 600] \quad T_c = [10 \quad 1000]$$

leading to a typical range for  $T_{br}$  to be [0.005 60]. This unrealistic range arises because the mathematics is uninformed about the correlation or functional dependency between  $T_b$  and  $T_c$ .

Multiple occurrences of variables does not always lead to excess width.  $T_{br}$  occurs twice in Equation 6.27.

$$h = T_{br} \frac{\ln P_c}{1 - T_{br}} \quad (6.27)$$

However, the excess width is zero. This is shown by inserting the interval  $T_{br} = [\underline{T}_{br} \ \ \overline{T}_{br}]$  into Equation 6.27 and using the definitions of interval arithmetic operations to determine the interval values for  $h$ . Inserting the interval definition of  $T_{br}$  into Equation 6.27 we start with

$$h = [\underline{T}_{br} \ \ \overline{T}_{br}] \frac{\ln \overline{T}_{br}}{1 - [\underline{T}_{br} \ \ \overline{T}_{br}]} \quad (6.28)$$

Using the interval arithmetic definitions for subtraction,

$$h = [\underline{T}_{br} \ \ \overline{T}_{br}] \frac{\ln P_c}{\left[1 - \overline{T}_{br} \ \ 1 - \underline{T}_{br}\right]} \quad (6.29)$$

division,

$$h = \left[ \underline{T}_{br} \ \ \overline{T}_{br} \right] \left[ \frac{\ln P_c}{1 - \underline{T}_{br}} \ \ \frac{\ln P_c}{1 - \overline{T}_{br}} \right] \quad (6.30)$$

and multiplication, noting that both intervals are greater than zero, we obtain the interval value of  $h$  to be

$$h = \left[ \frac{\underline{T}_{br} \ln P_c}{1 - \underline{T}_{br}} \ \ \frac{\overline{T}_{br} \ln P_c}{1 - \overline{T}_{br}} \right] \quad (6.31)$$

Since  $T_{br}$  is constrained to values between 0 and 1, Equation 6.28 is easily seen to be monotonic with respect to  $T_{br}$ . Therefore, the *exact* interval value for  $h$  can be

obtained by inserting the upper and lower interval bounds for  $T_{br}$  into the equation. This insertion produces an interval for  $h$  of

$$h = \left[ \frac{\underline{T}_{br} \ln P_c}{1 - \underline{T}_{br}} \quad \frac{\overline{T}_{br} \ln P_c}{1 - \overline{T}_{br}} \right] \quad (6.32)$$

which is identical to the interval using the interval arithmetic operation definitions.

For monotonic functions the interval extensions are equal to the function values at the end points. Thus for functions such as  $\sqrt{x}$ ,  $\text{expt } x$ , or  $\ln x$ , the interval extensions are simply

$$\vec{f}(X) = [f(\underline{X}) \quad f(\overline{X})] \quad (6.33)$$

It is important to note that taking advantage of the monotonicity of a function enables us to compute the interval value of the function exactly. Extending this concept of monotonicity of a function from simple functions to terms of an equation or even an entire equation I was able to modify physical property relationships so as to produce intervals of a much smaller width than those which would be produced using a naive interval evaluation. The identification of monotonic terms is detailed in Appendix D.

## 6.10 Meta-Algorithm

The generate and test algorithm described earlier must now be modified to deal with our abstractions. Instead of generating and testing individual molecules we generate and test meta-molecules. Those meta-molecules which satisfy the test are reduced in abstraction. This reduction in abstraction is accomplished by dividing one or more meta-groups. This produces a new generation of meta-molecules which are retested.

I demonstrate the algorithm using the meta-groups of Table 6.4 and the meta-contributions of Table 6.6. I design for the constraint:

$$T_b > 500 \text{ K} \quad (6.34)$$

Limiting the number of groups contained in a molecule between 2 and 4 we generate the following 65 meta-molecules:

(2 0 0 0)	(0 2 0 0)	(0 0 2 0)	(0 0 0 2)	(1 1 0 0)
(1 0 1 0)	(1 0 0 1)	(0 1 1 0)	(0 1 0 1)	(0 0 1 1)
(3 0 0 0)	(0 3 0 0)	(0 0 3 0)	(0 0 0 3)	(2 1 0 0)
(2 0 1 0)	(2 0 0 1)	(0 2 1 0)	(0 2 0 1)	(0 0 2 1)
(1 2 0 0)	(1 0 2 0)	(1 0 0 2)	(0 1 2 0)	(0 1 0 2)
(0 0 1 2)	(1 1 1 0)	(1 1 0 1)	(1 0 1 1)	(0 1 1 1)
(4 0 0 0)	(0 4 0 0)	(0 0 4 0)	(0 0 0 4)	(3 1 0 0)
(3 0 1 0)	(3 0 0 1)	(0 3 1 0)	(0 3 0 1)	(0 0 3 1)
(1 3 0 0)	(1 0 3 0)	(1 0 0 3)	(0 1 3 0)	(0 1 0 3)
(0 0 1 3)	(2 2 0 0)	(2 0 2 0)	(2 0 0 2)	(0 2 2 0)
(0 2 0 2)	(0 0 2 2)	(2 1 1 0)	(2 1 0 1)	(2 0 1 1)
(0 2 1 1)	(1 2 1 0)	(1 2 0 1)	(1 0 2 1)	(0 1 2 1)
(1 1 2 0)	(1 1 0 2)	(1 0 1 2)	(0 1 1 2)	(1 1 1 1)

Again the meta-molecule (1 0 1 2) represents the set of all molecules which can be formed by taking any one group from Meta-Group 1, no groups from Meta-Group 2, any one group from Meta-Group 3, and any two groups from Meta-Group 4.

Our 65 candidate molecules are then pruned using structural and property constraints. Chemical constraints were not used in the automatic design but kept until

the molecule screening step.

When we abstract groups into meta-groups we eliminate some of the information known about the group. For each of the meta-groups of Table 6.4 we know its ring class and global valence. All the meta-groups are acyclic. Meta-Group 1 has a global valence of one. Meta-Group 2 has a global valence of two. Meta-Group 3 has a global valence of three. Meta-Group 4 has a global valence of 4.

Knowing ring class and global valence we are able to use structural constraints 1, 2, 3, 4, 5, 6, 7, 9, and 10. Structural constraint 10 encompasses the others. It is the only one needed to be applied. The maximum valence any group has is four. Therefore, for each of our 65 meta-molecules we check to see that:

$$n_1 = 2 + n_3 + 2n_4 \quad (6.35)$$

Using this constraint we prune 61 meta-molecules. The remaining 4 meta-molecules are:

$$(2\ 0\ 0\ 0) \quad (2\ 1\ 0\ 0) \quad (3\ 0\ 1\ 0) \quad (2\ 2\ 0\ 0)$$

After applying our structural constraints we apply our property constraint. Table 6.7 shows  $T_b$  estimated for each of our four meta-molecules.  $T_b$  was estimated using the meta-contributions of Table 6.6 and Equation 6.16.

Applying our property constraint:

$$T_b > 500 \text{ K}$$

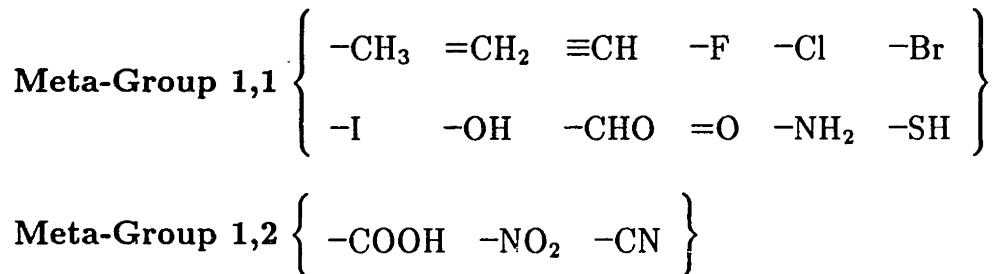
we see all four molecules contain values which satisfy the constraint.

Table 6.7:  $T_b$  Values for Four Meta-Molecules

Meta-Molecule	$T_b$
(2 0 0 0)	[177.18 536.36]
(2 1 0 0)	[199.60 617.46]
(3 0 1 0)	[178.42 729.59]
(2 2 0 0)	[222.02 698.56]

The next step of the meta-algorithm is to reduce the level of abstraction. We abstracted the groups into meta-groups to reduce the combinatorics of the problem. However, this same abstraction reduced the effectiveness of our property constraints. As we reduce the abstraction we regain this effectiveness.

I divide Meta-Group 1 into two new Meta-Groups:



This division of Meta-Group 1 is propagated to our meta-molecules. Meta-molecule (2 0 0 0) was the set of all molecules which could be formed by taking any two groups from Meta-Group 1. Now that Meta-Group 1 was divided into two new meta-groups we have three possibilities:

1. taking any two groups from Meta-Group 1,1
2. taking any two groups from Meta-Group 1,2
3. taking any one group from Meta-Group 1,1 and any one group from Meta-Group 1,2.

These possibilities correspond to meta-molecule (2 0 0 0) expanding into three new meta-molecules:

$$(2\ 0\ 0\ 0\ 0) \quad (0\ 2\ 0\ 0\ 0) \quad (1\ 1\ 0\ 0\ 0)$$

Expanding all four meta-molecules we obtain the following 13 new meta-molecules:

$$(2\ 0\ 0\ 0\ 0) \quad (0\ 2\ 0\ 0\ 0) \quad (1\ 1\ 0\ 0\ 0) \quad (2\ 0\ 1\ 0\ 0) \quad (0\ 2\ 1\ 0\ 0)$$

$$(1\ 1\ 1\ 0\ 0) \quad (3\ 0\ 0\ 1\ 0) \quad (0\ 3\ 0\ 1\ 0) \quad (2\ 1\ 0\ 1\ 0) \quad (1\ 2\ 0\ 1\ 0)$$

$$(2\ 0\ 2\ 0\ 0) \quad (0\ 2\ 2\ 0\ 0) \quad (1\ 1\ 2\ 0\ 0)$$

The meta-contributions toward  $T_b$  are also divided:

Group	$T_b$
Meta-Group 1,1	[−10.50 93.84]
Meta-Group 1,2	[125.66 169.09]
Meta-Group 2	[22.42 81.10]
Meta-Group 3	[11.74 24.14]
Meta-Group 4	[18.25 18.25]

We repeat the procedure estimating  $T_b$  for each of our meta-molecules and applying our property constraint. Table 6.8 shows estimated  $T_b$  values for the 13 meta-molecules.

Applying our property constraint prunes meta-molecules  $(2\ 0\ 0\ 0\ 0)$ ,  $(1\ 1\ 0\ 0\ 0)$ , and  $(2\ 0\ 1\ 0\ 0)$ . Additionally the meta-property for meta-molecule  $(0\ 3\ 0\ 1\ 0)$  shows that all the molecules it contains have  $T_b$  values which satisfy our property constraint. None of the meta-molecules resulting from further expansion of meta-molecule  $(0\ 3\ 0\ 1\ 0)$  need to be checked.

The meta-algorithm continues expanding meta-groups until all meta-groups contain only one group. At that point the abstraction has been removed and the meta-molecules generated represent individual molecules. The meta-algorithm can be summarized by the schematic shown in Figure 6.4.

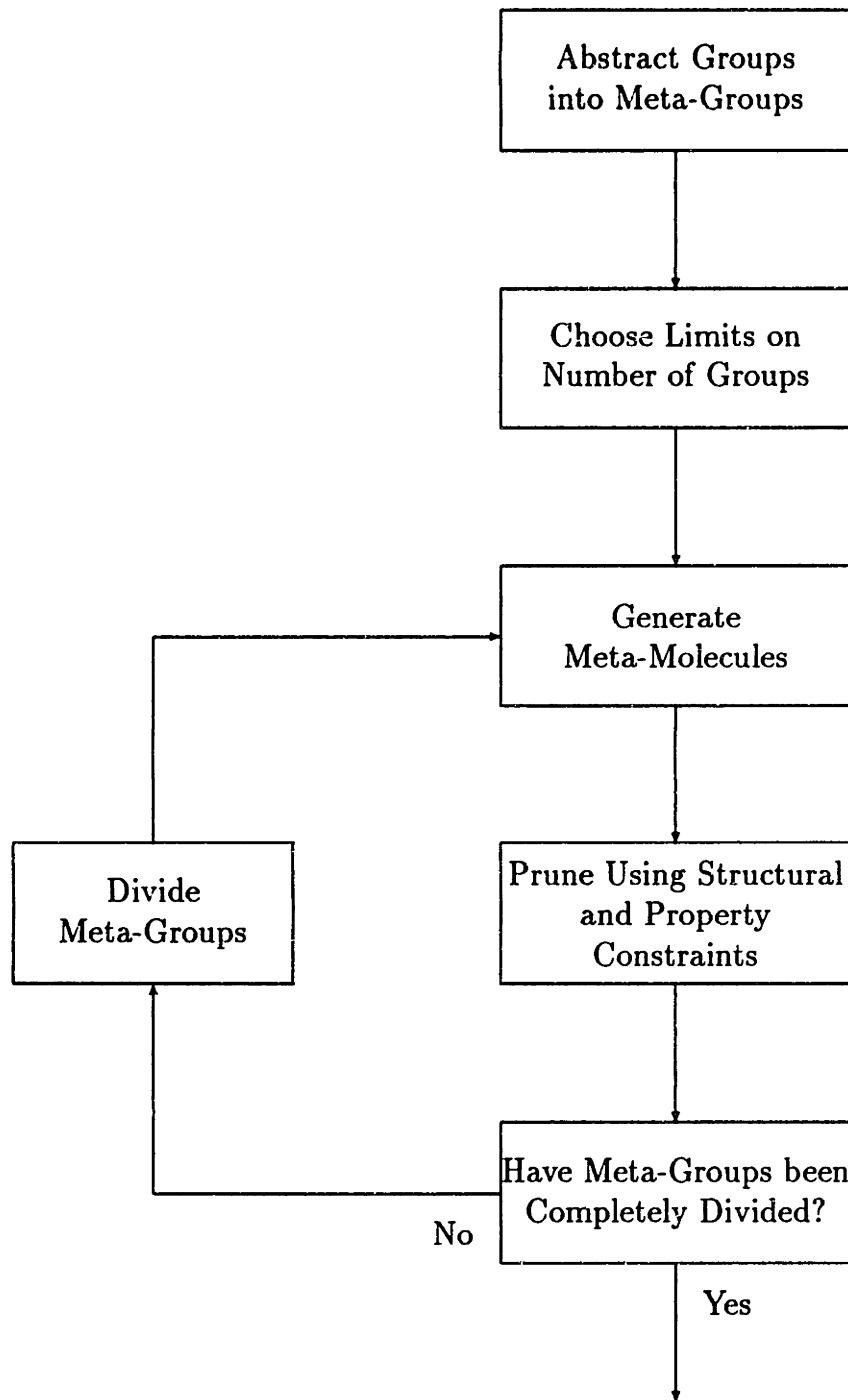


Figure 6.4: Automatic Design Meta-Algorithm

Table 6.8:  $T_b$  Values for 13 Meta-Molecules

(2 0 0 0 0)	[177.18 385.86]
(0 2 0 0 0)	[449.50 536.36]
(1 1 0 0 0)	[313.34 461.11]
(2 0 1 0 0)	[199.60 466.96]
(0 2 1 0 0)	[471.92 617.46]
(1 1 1 0 0)	[335.76 542.21]
(3 0 0 1 0)	[178.42 503.84]
(0 3 0 1 0)	[586.90 729.59]
(2 1 0 1 0)	[314.58 579.09]
(1 2 0 1 0)	[450.74 654.34]
(2 0 2 0 0)	[222.02 548.06]
(0 2 2 0 0)	[494.34 698.56]
(1 1 2 0 0)	[358.18 623.31]

One of the important issues in the meta-algorithm is the strategy used for dividing meta-groups.

### 6.10.1 Division Strategies

There are a large number of ways in which meta-group division can be done. Given a set of  $k$  objects the number of ways these can be partitioned into  $p$  sets is given by:

$$S(k,p)$$

which is the Stirling number of the second kind.

Stirling numbers of the second kind can be computed by noting that the number of ways of dividing  $k$  objects into 1 set,  $S(k,1)$  equals 1; the number of ways of dividing  $k$  objects into  $k$  sets,  $S(k,k)$  equals 1; and the recurrence relationship:

$$S(k,p) = S(k-1,p-1) + p S(k-1,p)$$

Table 6.9: Hypothetical Meta-Groups from Partitioning 4 Groups into 2 Clusters

$((a) (b c d))$	$((a b) (c d))$	$((a b c) (d))$
$((b) (a c d))$	$((b c) (a d))$	$((b d) (a c))$
$((a d) (b c))$		

Thus starting with a set of hypothetical groups:

$$(a b c d)$$

and dividing them into 2 meta-groups, we have  $S(4,2) = 7$  possibilities. These are shown in Table 6.9.

For a reasonable number of groups the possible choices of meta-groups is very large.

Table 6.10 shows the number of possible clusterings for an initial set of 19 groups.

Abstracting groups into meta-groups reduced the number of groups we had to deal with and hence the combinatorics. The reduction in detail also reduced the number and effectiveness of our structural and property constraints we use for pruning. Structural constraints require knowledge about the structural characteristics of groups: global valence, ring class, etc. Adding back detail involves separating our meta-groups into smaller collections. Forming these smaller collections can be done so that the new collections have a consistent set of characteristics or have a particularly narrow width of a meta-contribution toward a particular physical property.

I have termed these two approaches *expansion* and *division*. Expansion adds back knowledge about the meta-groups which is used by the structural constraints. Division focuses on reducing the width of meta-contributions thus improving the power of

Table 6.10: Partitioning 19 Groups into N Clusters

Number of Clusters	Number of Partitions
1	1
2	262,143
3	193,448,101
4	11,259,666,950
5	147,589,284,710
6	693,081,601,779
7	1,492,924,634,839
8	1,709,751,003,480
9	1,144,614,626,805
10	477,297,033,785
11	129,413,217,791
12	23,466,951,300
13	2,892,439,160
14	243,577,530
15	13,916,778
16	527,136
17	12,597
18	171
19	1

physical property constraints. I discuss various strategies for meta-group division.

**In Half** Dividing a meta-group in half is the simplest strategy. However, division without regard to the meta-contributions could prove inefficient. Given the following set of groups with contributions:

Group:	g1	g2	g3	g4
Contribution:	10	60	11	61

our initial meta-group  $\{g1\ g2\ g3\ g4\}$  would have a meta-contribution of  $[10\ 61]$ . Dividing the meta-group in half would result in the two meta-groups  $\{g1\ g2\}$  and  $\{g3\ g4\}$ . The meta-contributions for these new meta-groups would be  $[10\ 60]$  and  $[11\ 61]$ . These meta-contributions are almost identical to the original. The division thus added to the combinatorics without improving the possibility for pruning.

Meta-contributions should be considered when dividing meta-groups in half. The midpoint of our initial meta-contribution is:

$$\frac{61 - 10}{2} = 25.5$$

All groups whose contributions are less than 25.5 are collected into one new meta-group and all those with contributions greater than 25.5 are collected into a second new meta-group.

**Largest Gap** The interval representation for meta-contributions ignores their discrete nature. Dividing a meta-group at the largest gap in its contributions attempts to take advantage of this discrete nature to produce two new meta-groups whose meta-contributions are much narrower than the original meta-contribution. Using the same

example set of groups and contributions above, our initial meta-contribution, [10 61], has a width of 51. Dividing the meta-group at the largest gap in the contributions produces two new meta-groups with meta-contributions [10 11] and [60 61]. The total width of these two intervals is 2, a considerable reduction from 51.

**Isolating Groups** Extreme values for the contributions for some groups can greatly affect the value for calculated meta-properties. The contributions toward  $T_b$  from Joback's method[62] for =O is -10.5. If we are searching for low values of  $T_b$  then it is desirable to have many =O groups in our molecules. However, from chemical considerations it is unlikely that a molecule with a large number of =O groups would be stable. Isolating =O into its own meta-group enables constraints to be placed on the maximum occurrence of the meta-group in any meta-molecule.

**Meta-Group Occurrences** At times all meta-molecules which contain a certain meta-group are pruned away. The meta-group does not occur in any meta-molecule. Further division of the meta-group increase the combinatorics with no additional pruning possibilities.

## 6.11 Algorithm Evaluation

The number of variables involved in the automatic design make a detailed evaluation of the algorithm untenable. The performance of the algorithm is dependent upon the physical property constraints, chosen estimation procedures, the initial set of groups, and the range of group occurrences investigated. However, by making several assump-

tions it is possible to identify some of the “bounding” properties of the algorithm.

I begin by examining the combinatorics of the algorithm without considering any pruning. Without pruning the number of meta-molecules needing generation and testing is greater than the number of molecules contained in the search. This analysis shows the worst behavior of the algorithm.

Table 6.11 shows the number of possible meta-molecules for a 20 group, 3 occurrence automatic design. The total number of molecules contained in this design is 1540.

Table 6.12 shows the number of possible meta-molecules for a 20 group, 6 occurrence automatic design. The total number of molecules contained in this design is 177,100.

The meta-algorithm begins with all the groups abstracted into a single meta-group and ends with each meta-group containing only a single group. The tables show the maximum number of meta-molecules which could be present at any level of abstraction.

The number of meta-groups at any abstraction level is given by:

$$C^R(l, n) = C(l + n - 1, n) = \frac{(l + n - 1)!}{n!(l - 1)!} \quad (6.36)$$

where  $l$  is the abstraction level, equal to the number of meta-groups, and  $n$  is the occurrence value. The total number of meta-molecules generated is always greater than the molecules contained in the search if no constraints are considered.

The number of meta-molecules present at any level of abstraction does not correspond to the number of meta-molecules which need to be tested. Whenever the number of meta-groups is greater than one, meta-molecules have at least one zero occurrence. Table 6.13 shows the 10 meta-molecules formed from three meta-groups with an occurrence value of three.

Table 6.11: Maximum Number of Meta-Molecules

---

# groups = 20	# occurrences = 3
# Meta Groups	# Possible Meta-Molecules
1	1
2	4
3	10
4	20
5	35
6	56
7	84
8	120
9	165
10	220
11	286
12	364
13	455
14	560
15	680
16	816
17	969
18	1140
19	1330
20	1540
Total	8855

---

Table 6.12: Maximum Number of Meta-Molecules

---

# groups = 20	# occurrences = 6
# Meta Groups	# Possible Meta-Molecules
1	1
2	7
3	28
4	84
5	210
6	462
7	924
8	1,716
9	3,003
10	5,005
11	8,008
12	12,376
13	18,564
14	27,132
15	38,760
16	54,264
17	74,613
18	100,947
19	134,596
20	177,100
Total	657,800

---

Table 6.13: Meta-Molecules Formed from 3 Meta-Groups

---

(3 0 0)	(0 3 0)	(0 0 3)	(2 1 0)	(2 0 1)
(0 2 1)	(1 2 0)	(1 0 2)	(0 1 2)	(1 1 1)

---

We divide meta-groups to move to the next lower level of abstraction. Dividing a meta-group necessitates expanding all meta-molecules which contain occurrences of the meta-group. Meta-molecules (3 0 0), (0 3 0), (2 1 0), and (1 2 0) all contain zero occurrences of meta-group 3. Dividing meta-group 3 results in the four new meta-molecules (3 0 0 0), (0 3 0 0), (2 1 0 0), and (1 2 0 0). Although these meta-molecules have been expanded their meta-properties and feasibility are unchanged. The maximum number of meta-molecules which need to be generated and tested at any level is reduced by the number of meta-molecules which have zero in a specific meta-group location.

The number of meta-molecules having a zero in a specific location is given by fixing the  $l$ th location to zero and computing the combinatorics for the remaining  $l - 1$  locations. Using the example of three meta-groups with an occurrence value of three:

$$C^R(2, 3) = C(2 + 3 - 1, 3) = \frac{4!}{3! 1!} = 4 \quad (6.37)$$

There are 4 meta-molecules at the third level of abstraction which have zero occurrence of meta-group 3. At level  $l$  the number is given by:

$$C^R(l - 1, n) = C(l - 1 + n - 1, n) = \frac{(l + n - 2)!}{n!(l - 2)!} \quad (6.38)$$

The number of meta-molecules at any abstraction level greater than 2 needing testing is given by:

$$\frac{(l + n - 1)!}{n!(l - 1)!} - \frac{(l + n - 3)!}{n!(l - 3)!} \quad (6.39)$$

where  $l$  is the level of abstraction and  $n$  is the number of occurrences. Table 6.14 and Table 6.15 list the maximum number of meta-molecules which need to be tested at each abstraction level.

Table 6.14: Maximum Number of Meta-Molecules Needing Testing

# groups = 20	# occurrences = 3
# Meta Groups	# Possible Meta-Molecules
1	1
2	4
3	9
4	16
5	25
6	36
7	49
8	64
9	81
10	100
11	121
12	144
13	169
14	196
15	225
16	256
17	289
18	324
19	361
20	400
Total	2870

Table 6.15: Maximum Number of Meta-Molecules Needing Testing

---

# groups = 20	# occurrences = 6
# Meta Groups	# Possible Meta-Molecules
1	1
2	7
3	27
4	77
5	182
6	378
7	714
8	1254
9	2079
10	3289
11	5005
12	7371
13	10556
14	14756
15	20196
16	27132
17	35853
18	46683
19	59983
20	76153
Total	311696

---

The maximum number of meta-molecules which need to be tested in any automatic design is given by:

$$2 + n + \sum_{l=3}^k \left[ \frac{(l+n-1)!}{n!(l-1)!} - \frac{(l+n-3)!}{n!(l-3)!} \right] \quad (6.40)$$

By expanding Equation 6.40 out several terms we see that it simplifies to:

$$\frac{(k+n-2)!}{n!(k-2)!} + \frac{(k+n-1)!}{n!(k-1)!} \quad (6.41)$$

The total number of molecules in any automatic design is given by:

$$\frac{(k+n-1)!}{n!(k-1)!} \quad (6.42)$$

The ratio of the number of meta-molecules generated to the number of molecules contained in the search is given by:

$$\frac{\frac{(k+n-2)!}{n!(k-2)!} + \frac{(k+n-1)!}{n!(k-1)!}}{\frac{(k+n-1)!}{n!(k-1)!}} \quad (6.43)$$

This simplifies to:

$$\frac{k-1}{k+n-1} + 1. \quad (6.44)$$

For any  $n > 0$  the ratio is greater than 1.

Unless we consider the effect constraints have on pruning the search space the automatic design algorithm is seen to be less efficient than blind search. This is true of all algorithms which use a hierarchical abstraction of a search space. Unless we have constraints which can prune possibilities the process of abstraction is worthless.

Table 6.16 and Table 6.17 display two example automatic design runs. The design searched for molecules satisfying the constraint

$$P_{vp}(273K) > 1.0 \text{ bar} \quad (6.45)$$

Table 6.16: Pruning Results for k=44, n=3 Automatic Design

Constraint = $P_{vp}(273K) > 1.0$ bar.				
# Meta-Groups	# Meta-Molecules	Kept	Pruned	
1	1	1	0	
3 <sup>a</sup>	10	4	6	
4 <sup>b</sup>	7	4	3	
10 <sup>c</sup>	51	1	50	
11	3	1	2	
12 <sup>d</sup>	3	2	1	
13	5	3	2	
14	6	3	3	
15	6	4	2	
16	8	7	1	
17	11	11	0	
18	12	11	1	
19	21	18	3	
20	28	26	2	
21	33	29	4	
22	44	44	0	
27	109	85	24	
28 <sup>e</sup>	102	102	0	
Total <sup>f</sup> :	460	356	104	

<sup>a</sup>Expanded by Ring Class.

<sup>b</sup>Isolated -COOH, -NO<sub>2</sub>, and -CN.

<sup>c</sup>Expanded by global valence.

<sup>d</sup>Isolated =O. Restricted =O occurrences to 1.

<sup>e</sup>12 meta-groups never occurred in any meta-molecules.

<sup>f</sup>There are 15,180 molecules contained in the search

using an initial set of 44 groups. Table 6.18 and Table 6.19 show the percentage of meta-molecules pruned at each level of abstraction. Using these examples I continue my analysis with the assumption that a 10% average pruning at each level of abstraction is not unrealistic.

To account for an average value of pruning I need to establish an average value for the number of children generated from each meta-molecule. Table 6.20 shows the

Table 6.17: Pruning Results for k=44, n=5 Automatic Design

Constraint = $P_{vp}(273K) > 1.0$ bar.				
# Meta-Groups	# Meta-Molecules	Kept	Pruned	
1	1	1	0	
3 <sup>a</sup>	21	8	16	
4 <sup>b</sup>	23	8	15	
5 <sup>c</sup>	23	12	11	
6 <sup>d</sup>	30	27	3	
7	58	27	31	
8	58	32	26	
9	37	35	2	
10	53	35	18	
11	71	42	29	
12	87	86	1	
13	110	86	24	
14	165	107	58	
15	206	179	27	
24 <sup>e</sup>	1675	185	1490	
30 <sup>f</sup>	625	479	146	
37 <sup>g</sup>	888	688	200	
Total <sup>h</sup> :	4131	2037	2094	

<sup>a</sup>Expanded by Ring Class.

<sup>b</sup>Isolated  $\text{--COOH}$ ,  $\text{--NO}_2$ , and  $\text{--CN}$ .

<sup>c</sup>Isolated  $\text{=O}$ . Restricted  $\text{=O}$  occurrences to a maximum of 1.

<sup>d</sup>Isolated  $\text{--F}$ .

<sup>e</sup>Expanded by global valence.

<sup>f</sup>Expanded several meta-groups containing two or three groups in half.

<sup>g</sup>Expanded all nonzero occurring meta-groups to individual groups. 10 meta-groups never occurred in any meta-molecule.

<sup>h</sup>There are 1,712,304 molecules contained in the search.

Table 6.18: Example Pruning Percentage: 3 Occurrence

# Meta-Groups	% of Meta-Molecules Pruned
1	0.0
3	60.0
4	42.9
10	98.0
11	66.7
12	33.3
13	40.0
14	50.0
15	33.3
16	12.5
17	0.0
18	8.3
19	14.3
20	7.1
21	12.1
22	0.0
27	22.0
28	0.0
<hr/>	
Average:	27.8

Table 6.19: Example Pruning Percentage: 5 Occurrence

# Meta-Groups	% of Meta-Molecules Pruned
1	0.0
3	76.2
4	65.2
5	47.8
6	10.0
7	53.5
8	44.8
9	5.4
10	34.0
11	40.8
12	1.1
13	21.8
14	35.2
15	13.1
24	89.0
30	23.4
37	22.5
<hr/>	
Average:	34.3
<hr/>	

Table 6.20: Average Number of Children Meta-Molecules Needing Testing

# Meta Groups	Average # Children	
	3 Occurrence	6 Occurrence
1	4.00	7.00
2	2.25	3.86
3	1.78	2.85
4	1.56	2.36
5	1.44	2.08
6	1.36	1.89
7	1.31	1.76
8	1.27	1.66
9	1.23	1.58
10	1.21	1.52
11	1.19	1.47
12	1.17	1.43
13	1.16	1.40
14	1.15	1.37
15	1.14	1.34
16	1.13	1.32
17	1.12	1.30
18	1.11	1.28
19	1.11	1.27

average number of children meta-molecules generated in a 20 group automatic design.

Table 6.21 and Table 6.22 show the pruning results for a hypothetical automatic design assuming 10% of the meta-molecules are pruned at each abstraction level excluding the first. The total number of meta-molecules searched is less than the number of molecules a blind search would examine.

One of the major factors in reducing the number of meta-molecules needing to be generated and tested is the identification of meta-groups which have been excluded from all meta-molecules. In the example automatic design shown in Table 6.16 I isolated the three groups  $\text{--COOH}$ ,  $\text{--NO}_2$ , and  $\text{--CN}$  into a separate meta-group. This was done

Table 6.21: Automatic Design with 10% Average Pruning

Meta-Groups	Meta-Molecules	Kept	Pruned
1	1.00	1.00	0.00
2	4.00	3.60	0.40
3	8.10	7.29	0.81
4	12.96	11.66	1.30
5	18.23	16.40	1.82
6	23.62	21.26	2.36
7	28.93	26.04	2.89
8	34.01	30.61	3.40
9	38.74	34.87	3.87
10	43.05	38.74	4.30
11	46.88	42.19	4.69
12	50.21	45.19	5.02
13	53.03	47.73	5.30
14	55.36	49.82	5.54
15	57.19	51.47	5.72
16	58.56	52.71	5.86
17	59.50	53.55	5.95
18	60.04	54.03	6.00
19	60.20	54.18	6.02
20	60.04	54.03	6.00
Total:	773.65	696.37	77.25

Table 6.22: Automatic Design with 10% Average Pruning

Meta-Groups	Meta-Molecules	Kept	Pruned
1	1.00	1.00	0.00
2	7.00	6.30	0.70
3	24.30	21.87	2.43
4	62.37	56.13	6.24
5	132.68	119.41	13.27
6	248.01	223.21	24.80
7	421.61	379.45	42.16
8	666.43	599.78	66.64
9	994.38	894.94	99.44
10	1415.81	1274.23	141.58
11	1939.04	1745.14	193.90
12	2570.11	2313.10	257.01
13	3312.58	2981.33	331.26
14	4167.53	3750.78	416.75
15	5133.55	4620.20	513.36
16	6206.93	5586.24	620.69
17	7381.81	6643.63	738.18
18	8650.45	7785.41	865.05
19	10003.47	9003.13	1000.35
20	11430.16	10287.14	1143.02
Total:	64769.22	58292.42	6476.83

because it was noticed in previous designs that including these groups in a molecule resulted in that molecule having a very low vapor pressure. It was subsequently found in that design that all molecules which contained occurrences of the meta-group containing the three isolated groups were pruned.

To examine this behavior I investigated the following scenario. The initial meta-group is divided into two children meta-groups. The meta-molecules formed from these groups are tested and all those which contain occurrences of the second group are pruned away. This scenario is repeated with the surviving meta-group and so on. This is favorable pruning and suggests a favorable bound on the algorithm.

Dividing a meta-group, MG, containing  $k$  groups into two meta-groups,  $MG_1$  and  $MG_2$ , containing  $k_1$  and  $k_2$  groups respectively allocates the

$$\frac{(k+n-1)!}{n!(k-1)!} \quad (6.46)$$

possible molecules into

$$\frac{(2+n-1)!}{n!(2-1)!} = n+1 \quad (6.47)$$

meta-molecules. One meta-molecule contains only occurrences of  $MG_1$ , one meta-molecule contains only occurrences of  $MG_2$ , and the remaining  $n-1$  meta-molecules contain occurrences of both meta-groups. If no meta-molecules containing  $MG_2$  survive the testing then the percentage of molecules pruned is given by

$$\left[ 1 - \frac{(k-1)!}{(k_1-1)!} \frac{(k_1+n-1)!}{(k+n-1)!} \right] \times 100\% \quad (6.48)$$

Repeating this expansion and pruning process  $r$  times until the final meta-group contains only one group requires the generation and testing of

$$r(n+1) + 1 \quad (6.49)$$

Table 6.23: Advantage of Abstraction:  $MG_2$  Contains 1 Group

$k \setminus n$	2	3	4	5	6
15	5.6	11.9	43.1	136.8	391.5
20	7.2	20.0	92.2	369.6	1,321.6
25	8.9	30.2	169.2	819.0	3,513.5
30	10.6	42.4	280.3	1,590.0	7,956.7
35	12.2	56.7	431.7	2,808.6	16,060.2
40	13.9	73.1	629.6	4,621.3	29,726.5
45	15.6	91.6	880.5	7,195.8	51,426.2
50	17.2	112.2	1,190.3	10,720.4	84,272.3

meta-molecules. The advantage of abstraction, as measured by the total number of molecules contained in the search divided by the number of meta-molecules needed to be generated and tested, is given by:

$$\text{Advantage of Abstraction} = \frac{1}{r(n+1)+1} \frac{(k+n-1)!}{n!(k-1)!} \quad (6.50)$$

Considering the worst case in which  $MG_2$  and all subsequent second meta-groups contain only a single group we have  $r = k - 1$  leading to:

$$\text{Advantage of Abstraction} = \frac{1}{(k-1)(n+1)+1} \frac{(k+n-1)!}{n!(k-1)!} \quad (6.51)$$

Table 6.23 shows this advantage of abstraction for several values of  $k$  and  $n$ .

Table 6.23 shows that if we have an automatic design involving 40 groups with an occurrence value of 5 then the number of meta-molecules needed for exhaustive searching is 4,621.3 times fewer than the number of molecules.

# Chapter 7

## Interactive Design

The interactive design technique uses the designer's knowledge to guide the search for new molecules. This knowledge could be:

1. which physical property constraints are hard and which are soft. Minor violations of soft constraints are acceptable.
2. atoms or groups the new molecule must contain or must not contain.
3. existing compounds which might be structurally modified to provide desirable properties.

Representing and manipulating this knowledge in the automatic design procedure is difficult. The interactive design procedure's representation of the design problem easily incorporates this knowledge.

I begin this chapter by describing the graphical representation used by the interactive design procedure. I show how the representation can accommodate a variety of estimation techniques enabling design for a fair number of physical properties. I finally show how the graphical representation of the design procedure provides significant insight into the molecular design problem.

Table 7.1: Example of Linear Group Contribution Estimation Techniques

Groups	Contributions	
	$\Delta_{i,T_b}$	$\Delta_{i,T_m}$
$-\text{CH}_3$	23.58	-5.10
$-\text{CH}_2-$	22.88	11.27
$>\text{CH}-$	21.74	12.64
$>\text{C}<$	18.25	46.43
-F	-0.03	-15.78
-Cl	38.13	13.55

## 7.1 Procedure Basis

Like the automatic design procedure, the interactive design procedure is based upon group contribution estimation techniques. The interactive design procedure takes advantage of the linearity of many group contribution techniques. One example of a linear technique is Joback's method[62] for estimating the normal boiling point,  $T_b$ . Table 7.1 shows the linear estimation model and an example set of group contributions.

The linear model enables us to assemble our molecule group by group. Given a constraint, such as

$$T_b > 300 \text{ K}$$

we choose a group for addition to our molecule and then evaluate the constraint. Choosing a  $-\text{CH}_3$  group we estimate  $T_b$  to be 221.76 K. Our constraint is not satisfied. Choosing a second  $-\text{CH}_3$  we estimate  $T_b$  to be 245.34. Our constraint is still not satisfied. Continuing our selection we find that after choosing five  $-\text{CH}_3$  groups we estimate a  $T_b$  of 316.08 which satisfies our constraint.

Although the five  $-\text{CH}_3$  groups satisfy our physical property target, they do not

form a structurally feasible molecule. Unlike the automatic design, the interactive design leaves the task of ensuring structural feasibility up to the designer.

To assist in selecting groups which satisfy our physical property constraints and are structurally feasible the following heuristics on group selection are often helpful:

- Separate the groups into three sets:

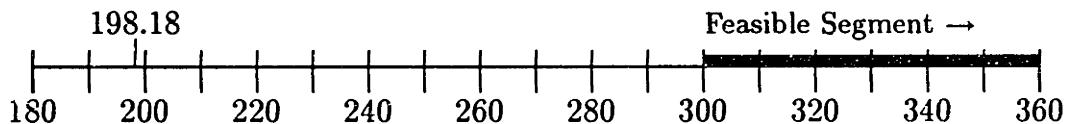
1. terminators: all groups having one free bond.
2. extenders: all groups having two free bonds.
3. branchers: all groups having more than two free bonds.

- Select the following initial group sets:

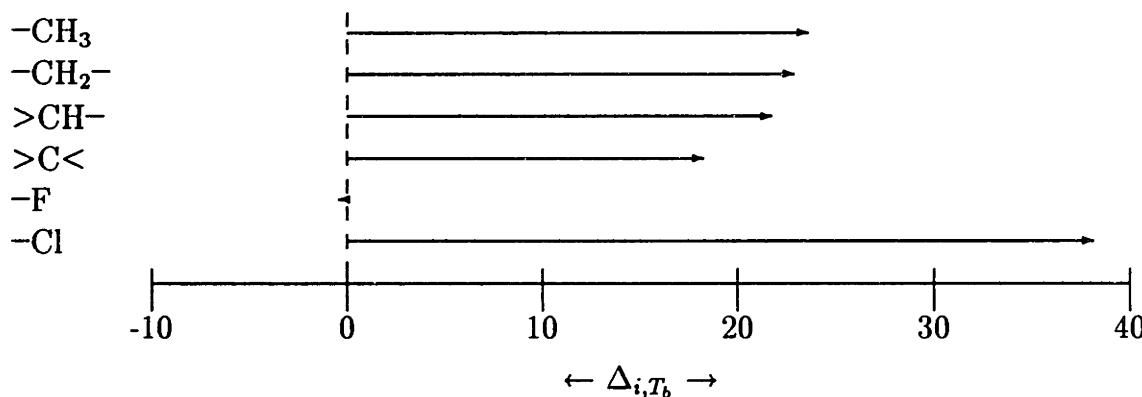
1. if a pure acyclic molecule is to be designed choose two terminators. If two terminators can not be found which satisfy our constraints then examine the addition of extenders. Extenders can be added without affecting the candidate's structural feasibility. If branchers must be added then follow immediately with the addition of terminators.
2. if a pure cyclic molecule is to be designed for then choose two cyclic extenders.

The theme of the heuristics is to ensure we have a structurally feasible molecule during most of our interactive design. Selecting 4 or 5 acyclic branchers giving a partial design satisfying our constraints is unwise because at least 5 or 6 terminators will also need to be chosen before we have a structurally feasible molecule.

Graphical representation of the design problem facilitates selecting groups. The normal boiling point is represented on a number line with our constraint denoting a feasible segment:

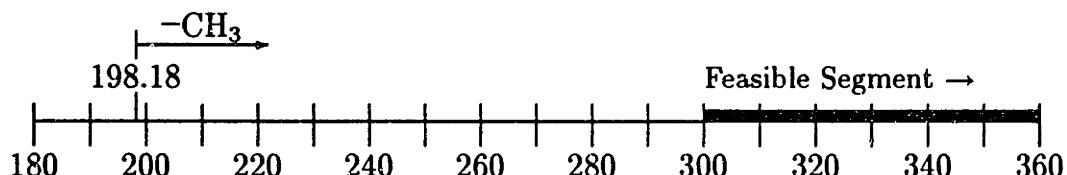


The intercept of our estimation technique is 198.18 and is noted on the number line. The contributions toward  $T_b$  from Table 7.1 are represented as vectors:

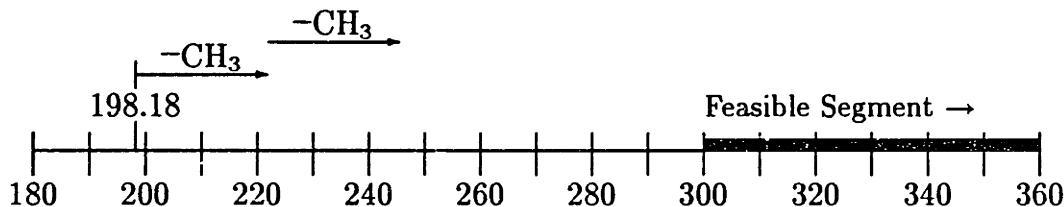


I call these vectors *group vectors*. The graphical representation of the contributions as vectors makes it easier to grasp the relative impact of choosing one group over another.

For example choosing our first  $-\text{CH}_3$  group corresponds graphically to adding a  $-\text{CH}_3$  group vector starting at the  $T_b$  intercept



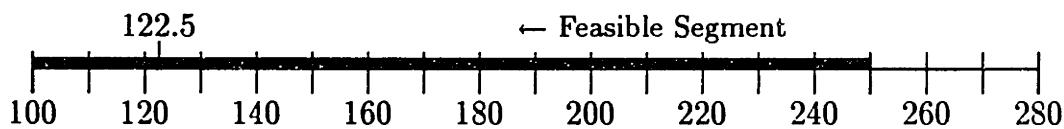
Adding a second  $-\text{CH}_3$  corresponds graphically to a  $-\text{CH}_3$  group vector beginning at the head of the current group vector



The representation shows its true advantage when dealing with a two dimensional example. Suppose we add a second constraint:

$$T_m < 250\text{K}$$

The normal melting point is estimated by Joback's estimation technique[62]. An example set of the contributions and the techniques model are shown in Table 7.1. In an analogous manner to the normal boiling point we form a number line on which to design for  $T_m$



and represent the group contributions as vectors

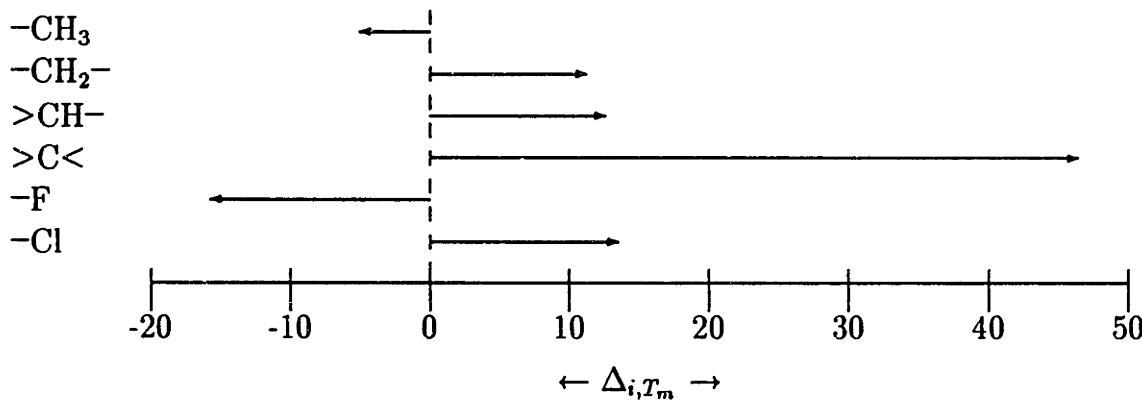


Figure 7.1 shows the two number lines combined into a 2-dimensional design space. The feasible segment of the number line now expands into a feasible region. This is denoted by the shaded area. The contributions for each group forms a 2-dimensional

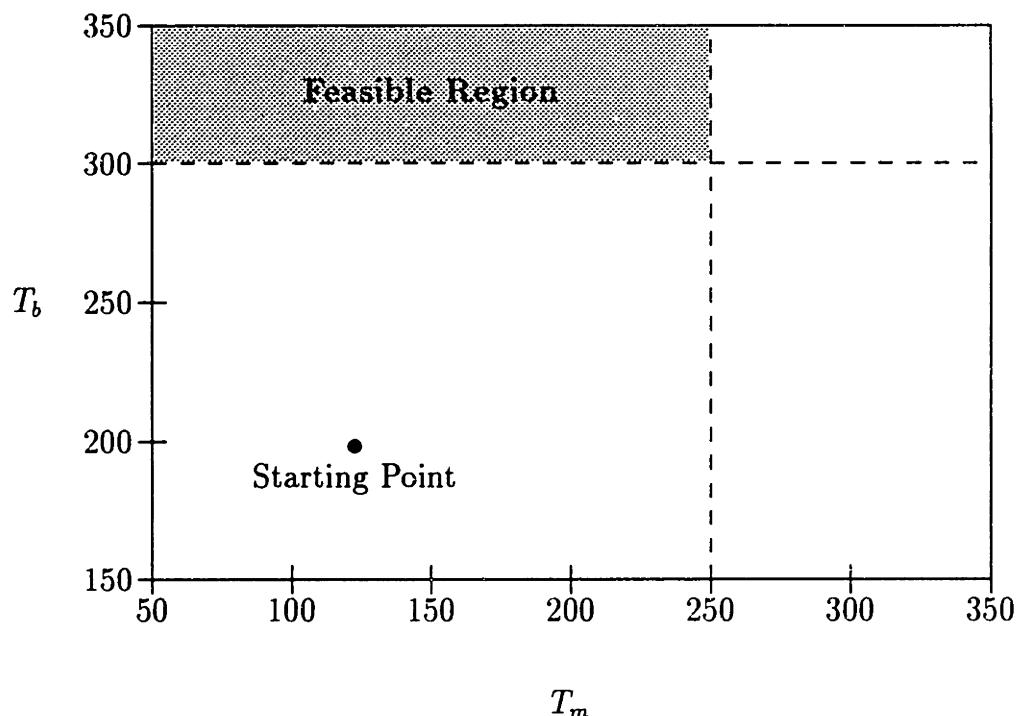


Figure 7.1: Graphical Representation of Constraints in a Physical Property Space

vector. Figure 7.2 shows our example group vectors.

As in the one dimensional example we begin at the intercept point selecting an appropriate set of group vectors which bring us into the feasible region and gives us a structurally feasible molecule. Figure 7.3 shows the group vectors for chloropropane.

Complex constraints on  $T_b$  and  $T_m$  are easily handled by the interactive design procedure. All that is required is to identify a feasible region or regions. The constraints need not be linear or convex.

Adding a third constraint to be designed for can be handled in two ways. Let this third constraint be:

$$\Delta H_{vb} > 5500 \text{ cal/g-mol} \quad (7.1)$$

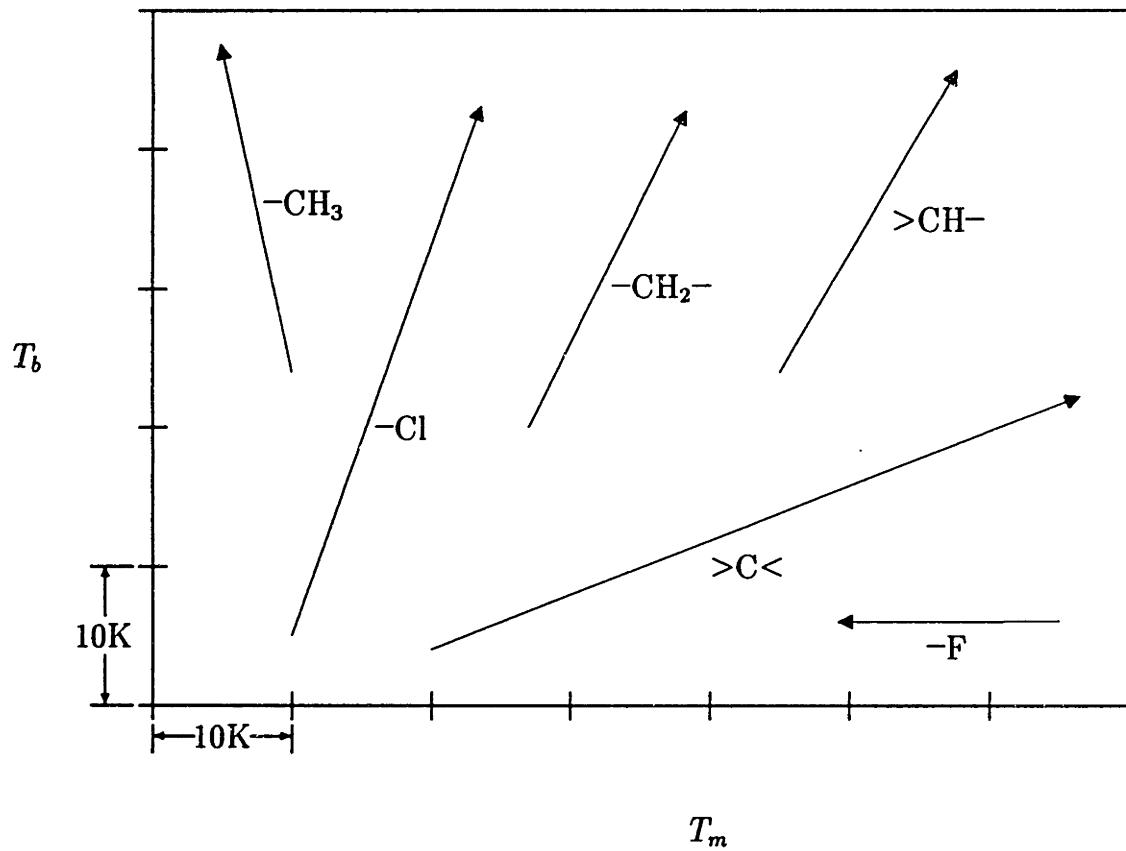


Figure 7.2: Graphical Representation of  $T_b$  and  $T_m$  Contributions

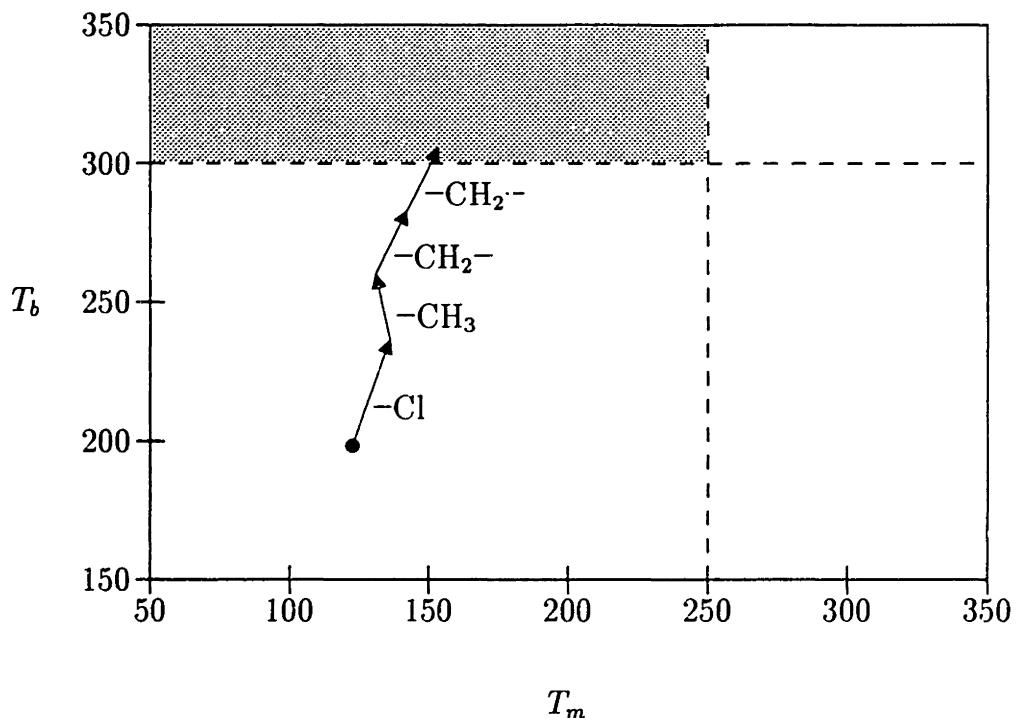


Figure 7.3: An Interactively Designed Molecule: Chloropropane

To estimate  $\Delta H_{vb}$  I use Joback's estimation method[62]. Table 7.2 displays the estimation model and the group contributions for our example groups.

Table 7.2:  $\Delta H_{vb}$  Group Contribution Estimation Technique

Groups	Contributions
	$\Delta_{i,\Delta H_{vb}}$
$-\text{CH}_3$	567
$-\text{CH}_2-$	532
$>\text{CH}-$	404
$>\text{C}<$	152
$-\text{F}$	-160
$-\text{Cl}$	1083

$$\Delta H_{vb} = 3657 + \sum n_i \Delta_{i,\Delta H_{vb}}$$

The first approach to including this third constraint is simply to add another axis to the design space and to perform our design in a 3-dimensional space. This leads to the difficulty of visualizing vectors in a three dimensional space. A second approach is to remember that the  $T_b$ - $T_m$  design space was constructed for ease of visualization not functional dependency. Thus we could also design simultaneously in two 2-dimensional design spaces:

$$T_b \text{ vs. } T_m \quad \text{and} \quad T_b \text{ vs. } \Delta H_{vb}.$$

Figure 7.4 shows 1,1,1-dichlorofluoroethane satisfying all three constraints.

A number of estimation techniques modify their estimated property to improve accuracy. Joback's method[62] for estimating the critical pressure and critical temperature are two examples. The estimation models for these two properties are:

$$T_{br} = \frac{T_b}{T_c} = 0.584 + 0.965 \sum n_i \Delta_{i,T_{br}} - \left( \sum n_i \Delta_{i,T_{br}} \right)^2 \quad (7.2)$$

$$P_c = \left( 0.113 + \sum n_i \Delta_{i,P_c} \right)^{-2} \quad (7.3)$$

To interactively design for these properties we must linearize them.

The model for  $T_{br}$  is linearized by solving the quadratic to yield:

$$\sum n_i \Delta_{i,T_{br}} = \frac{0.965 \pm \sqrt{(0.965)^2 - 4(T_{br} - 0.584)}}{2} \quad (7.4)$$

The negative root was chosen to be consistent with the given group contribution values.

Equation 7.4 expands to give:

$$\sum n_i \Delta_{i,T_{br}} = 0.4825 - \sqrt{0.817 - T_{br}} \quad (7.5)$$

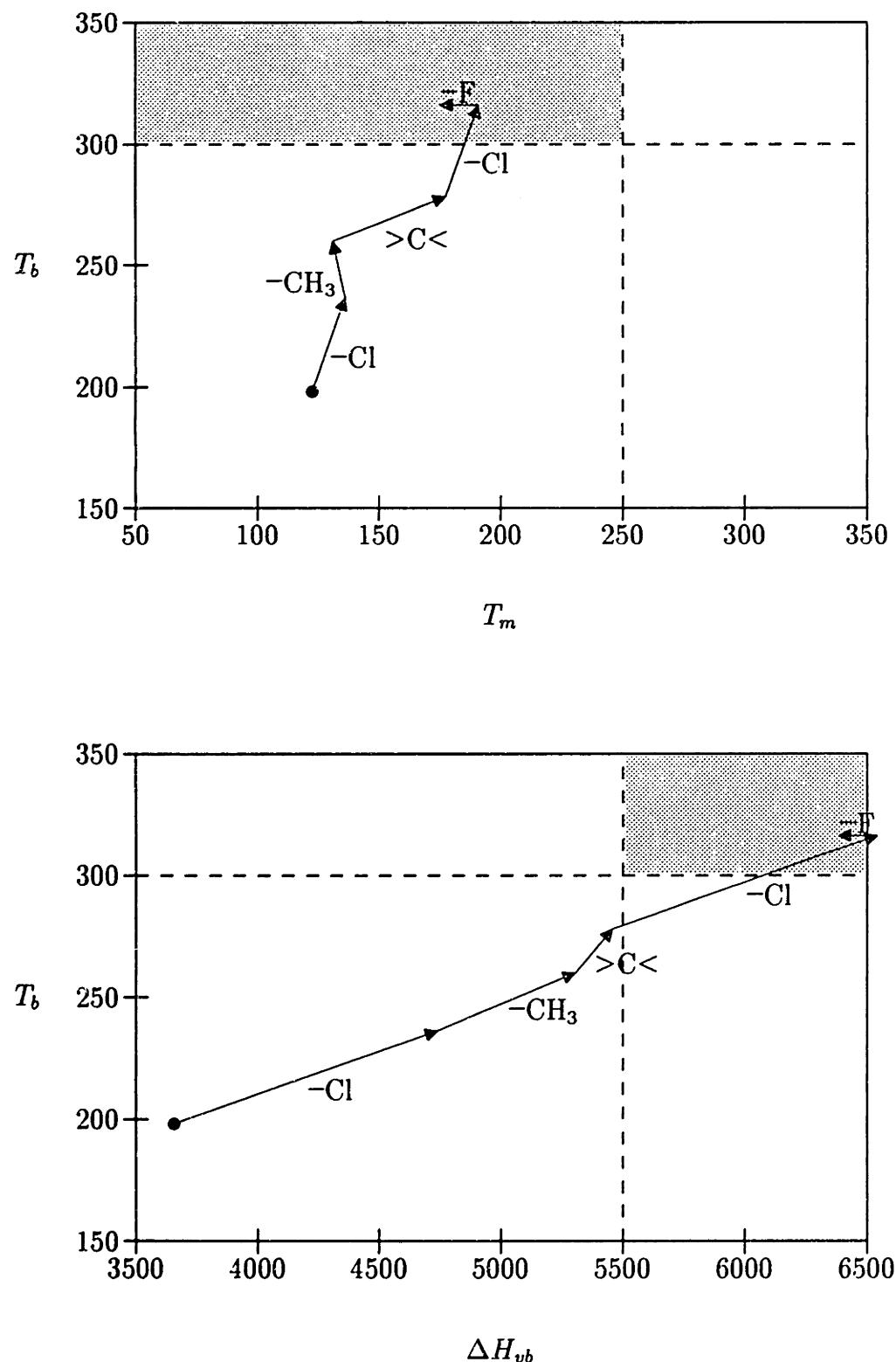


Figure 7.4: Simultaneous Design in Multiple Physical Property Spaces

rearranging

$$\sqrt{0.817 - T_{br}} = 0.4825 - \sum n_i \Delta_{i,T_{br}} \quad (7.6)$$

I consider the left hand side of Equation 7.6 to be a new property,  $T_{br}^*$ . Since it is linearly estimated, we are able to design for it in a physical property space.

Similarly defining:

$$P_c^* = \sqrt{\frac{1}{P_c}} \quad (7.7)$$

we have a linear model which can be used in a physical property space.

I term group contribution techniques which estimate a function of a physical property *combined estimation techniques*. The name comes from the fact that the technique is actually a combination of a group contribution technique and an equation oriented technique. For example, the combined technique for the estimation of  $P_c$  is made from the group contribution technique having the model:

$$P_c^* = 0.113 + \sum n_i \Delta_{i,P_c^*} \quad (7.8)$$

and the equation oriented technique

$$P_c = \frac{1}{(P_c^*)^2} \quad (7.9)$$

When performing an interactive design on constraints which are estimated by combined techniques the constraints must be propagated to constraints on the linearly estimated property. In designing for the constraints

$$T_{br} > 0.75 \quad (7.10)$$

$$P_c < 40 \text{ bar} \quad (7.11)$$

we propagate these to constraints on  $T_{b_r}^*$  and  $P_c^*$ :

$$T_{b_r}^* < 0.259 \quad (7.12)$$

$$P_c^* > 0.158 \text{ bar}^{-1/2} \quad (7.13)$$

These constraints are then plotted in a  $T_{b_r}^* - P_c^*$  physical property space and the design proceeds as before.

Combined estimation techniques extend the properties which can be designed for beyond those for which group contribution estimation techniques are available. Similarly, equation oriented techniques propagate constraints on the estimated property to fundamental properties. The acentric factor,  $\omega$ , is adequately estimated by equation oriented techniques but not by group contribution techniques. To design for the constraint

$$\omega > 0.65 \quad (7.14)$$

we use the equation oriented estimation technique[72]

$$\omega = \frac{-\ln P_c - 5.92714 + 6.09648/T_{b_r} + 1.28862 \ln T_{b_r} - 0.169347 T_{b_r}^6}{15.2518 - 15.6875/T_{b_r} - 13.4721 \ln T_{b_r} + 0.43577 T_{b_r}^6} \quad (7.15)$$

to propagate this constraint to one on  $T_{b_r}$  and  $P_c$ :

$$\frac{-\ln P_c - 5.92714 + 6.09648/T_{b_r} + 1.28862 \ln T_{b_r} - 0.169347 T_{b_r}^6}{15.2518 - 15.6875/T_{b_r} - 13.4721 \ln T_{b_r} + 0.43577 T_{b_r}^6} > 0.65 \quad (7.16)$$

Both  $T_{b_r}$  and  $P_c$  can be used in the interactive design procedure. Creating a design space from the linearized properties  $T_{b_r}^*$  and  $P_c$  we can plot Equation 7.16. Figure 7.5 shows the design space with its feasible region.

The equation oriented estimation technique for  $\omega$  allowed us to propagate constraints on  $\omega$  to constraints on  $T_{b_r}^*$  and  $P_c^*$ . Two fundamental physical properties re-

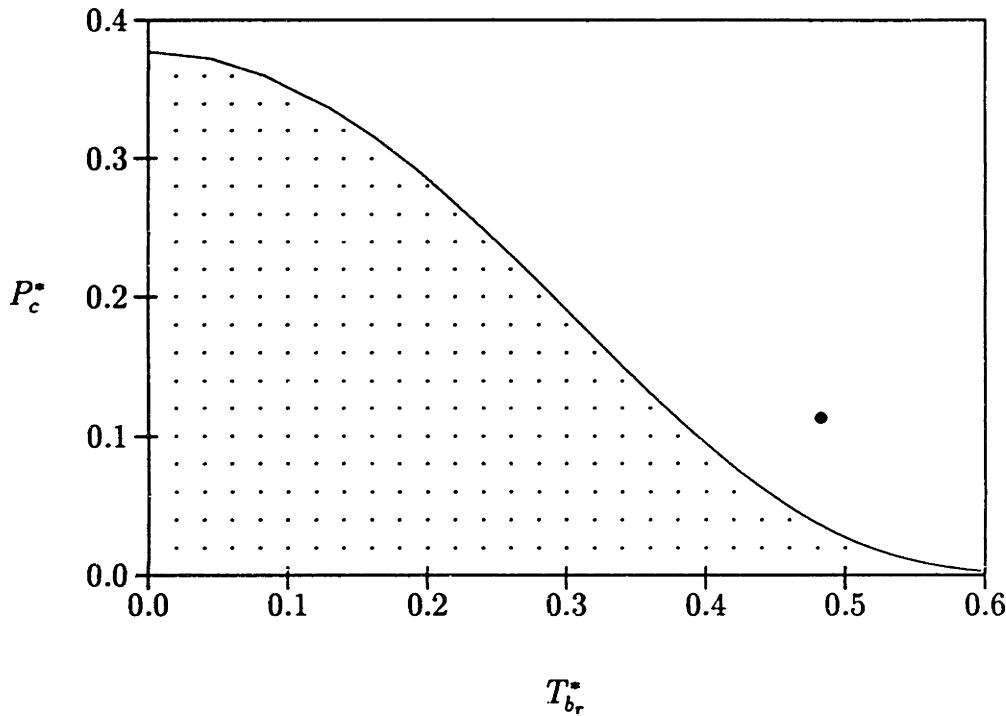


Figure 7.5:  $P_c^*$ - $T_{br}^*$  Design Space for Acentric Factor Constraint

sults in a two dimensional design space. Other equation oriented estimation techniques require more than two fundamental properties.

One estimation procedure for  $P_{vp}$  is:

- 1)  $P_{vp} = P_{vp}(T_b, T_{br}, P_c)$  by Riedel Plank Miller EOT
- 2)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(T_{br}^*)$  by Joback  $T_{br}$  Modification
- 2)  $P_c = P_c(P_c^*)$  by Joback  $P_c$  Modification
- 3)  $T_{br}^* = T_{br}^*(\text{groups})$  by Joback  $T_{br}^*$  GCT
- 3)  $P_c^* = P_c^*(\text{groups})$  by Joback  $P_c^*$  GCT

The fundamental properties are:  $T_b$ ,  $T_{br}^*$ , and  $P_c^*$ . Designing for constraints on  $P_{vp}$  requires a 3-dimensional design space. The physical property space containing our  $P_{vp}$  constraint is not separable as was our original design space shown in Figure 7.1.

This is because the constraint is a function of all the fundamental properties. This is unfortunate because representing and manipulating 3D objects graphically is more complex than 2D manipulation.

The dimensionality of our design space can become more than a mere complication.

One estimation procedure for the liquid heat capacity is:

- 1)  $C_{pL} = C_{pL}(\omega, C_p^o, T_c)$  by Rowlinson EOT
- 2)  $\omega = \omega(T_{br}, P_c)$  by Lee-Kesler EOT
- 2)  $T_c = T_c(T_b, T_{br})$  by definition
- 2)  $C_p^o = C_p^o(C_{p,a}^o, C_{p,b}^o, C_{p,c}^o, C_{p,d}^o)$  by  $C_p^o$  cubic fit
- 3)  $T_{br} = T_{br}(T_{br}^*)$  by Joback  $T_{br}$  Modification
- 3)  $P_c = P_c(P_c^*)$  by Joback  $P_c$  Modification
  - 4)  $T_{br}^* = T_{br}^*(\text{groups})$  by Joback  $T_{br}^*$  GCT
  - 4)  $P_c^* = P_c^*(\text{groups})$  by Joback  $P_c^*$  GCT
- 3)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
- 3)  $C_{p,a}^o = C_{p,a}^o(\text{groups})$  by Joback  $C_{p,a}^o$  GCT
- 3)  $C_{p,b}^o = C_{p,b}^o(\text{groups})$  by Joback  $C_{p,b}^o$  GCT
- 3)  $C_{p,c}^o = C_{p,c}^o(\text{groups})$  by Joback  $C_{p,c}^o$  GCT
- 3)  $C_{p,d}^o = C_{p,d}^o(\text{groups})$  by Joback  $C_{p,d}^o$  GCT

The fundamental properties are:  $T_{br}^*$ ,  $P_c^*$ ,  $T_b$ ,  $C_{p,a}^o$ ,  $C_{p,b}^o$ ,  $C_{p,c}^o$ ,  $C_{p,d}^o$ . To interactively design for constraints on  $C_{pL}$  would require a 7 dimensional physical property space to display each of the seven fundamental properties.

Factor analytic studies[20,21,61,67] show that a number of physical properties are highly intercorrelated. These studies are described in Appendix E. High correlations between two properties indicate the possibility of replacing one with a function of the other. This would enable us to reduce the dimensionality of our design space.

Table 7.3: Equations Relating Physical Properties to Factors

---

$1/\sqrt{P_c}$	=	0.157	—	0.019 $F_1$
$T_c$	=	545.9	—	24.65 $F_1$ — 87.92 $F_3$
$T_b$	=	358.4	—	25.26 $F_1$ — 64.94 $F_3$
$\Delta H_{vb}$	=	7686.6	—	432.3 $F_1$ — 1614.3 $F_3$

---

In one study[61] it was found that 9 physical properties were well approximated by 3 new properties called factors. Table 7.3 shows four of the derived equation oriented estimation techniques. Joback[61] developed group contributions for  $F_1$  and  $F_3$ .

We can incorporate these equation oriented estimation into a new estimation procedure for  $P_{vp}$

- 1)  $P_{vp} = P_{vp}(T_b, T_c, P_c)$  by Riedel Plank Miller EOT
- 2)  $T_b = T_b(F_1, F_3)$  by  $T_b$  Factor EOT
- 2)  $T_c = T_c(F_1, F_3)$  by  $T_c$  Factor EOT
- 2)  $P_c = P_c(F_1, F_3)$  by  $P_c$  Factor EOT
- 3)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
- 3)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

The fundamental properties are  $F_1$  and  $F_3$ . Two fundamental properties enable us to design in a two dimensional physical property space.

## 7.2 Constraint Visualization

Transformations made to reduce the design space to two dimensions enable us to visualize the tradeoffs between constraints. For example, using the two estimation procedures

- 1)  $P_{vp} = P_{vp}(T_b, T_c, P_c)$  by Riedel Plank Miller EOT

- 2)  $T_b = T_b(F_1, F_3)$  by  $T_b$  Factor EOT
- 2)  $T_c = T_c(F_1, F_3)$  by  $T_c$  Factor EOT
- 2)  $P_c = P_c(F_1, F_3)$  by  $P_c$  Factor EOT
- 3)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
- 3)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

and

- 1)  $\Delta H_v = \Delta H_v(\Delta H_{vb}, T_b, T_c)$  by Watson EOT
- 2)  $\Delta H_{vb} = \Delta H_{vb}(F_1, F_3)$  by  $\Delta H_{vb}$  Factor EOT
- 2)  $T_b = T_b(F_1, F_3)$  by  $T_b$  Factor EOT
- 2)  $T_c = T_c(F_1, F_3)$  by  $T_c$  Factor EOT
- 3)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
- 3)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

we are able to design for the constraints

$$P_{vp}(316.25) > 14 \text{ bar} \quad (7.17)$$

$$\Delta H_v(272.05) > 18.4 \text{ cal/mol} \quad (7.18)$$

in the two dimensional space shown in Figure 7.6. For the majority of the range for typical molecules the enthalpy of vaporization constraint is redundant. Being able to visualize the constraint provides insight into the relative tradeoffs of constraints.

Visualizing the tradeoffs among constraints often leads to investigating the sensitivity of the feasible region. Figure 7.6 showed a  $F_1$ - $F_3$  design space. The shaded area denotes the feasible region formed by the four constraints:

$$P_{vp}(272.05 K) > 1.4 \text{ bar} \quad (7.19)$$

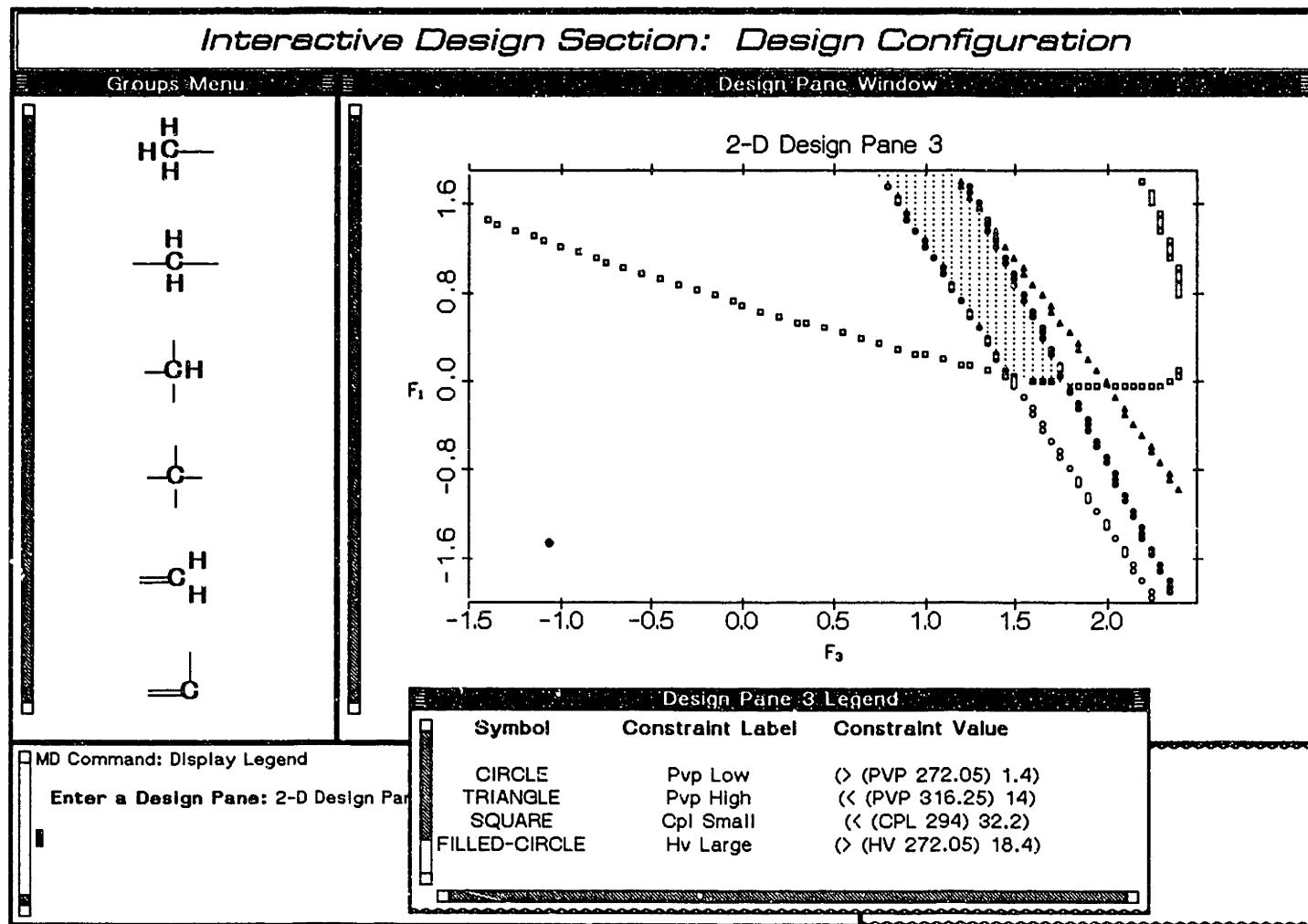


Figure 7.6: Two Dimensional Space Showing Four Constraints

$$P_{vp}(316.25K) < 14 \text{ bar} \quad (7.20)$$

$$C_{p_L}(294K) < 32.2 \text{ cal/mol}\cdot\text{K} \quad (7.21)$$

$$\Delta H_v(272.05K) > 18.4 \text{ kJ/mol} \quad (7.22)$$

Investigating the sensitivity of the feasible region to the  $\Delta H_v$  constraint we modify the constraint to:

$$\Delta H_v(272.05K) > 25 \text{ kJ/mol} \quad (7.23)$$

Figure 7.7 shows that this modification eliminates the feasible region for typical values of  $F_1$  and  $F_2$ .

Likewise if we considered the feasible region too restrictive we could “loosen” each of the constraints. Figure 7.8 shows the expanded feasible region formed by relaxing our constraints by 20% to:

$$P_{vp}(272.05K) > 1.12 \text{ bar} \quad (7.24)$$

$$P_{vp}(316.25K) < 16.8 \text{ bar} \quad (7.25)$$

$$C_{p_L}(294K) < 38.64 \text{ cal/mol}\cdot\text{K} \quad (7.26)$$

$$\Delta H_v(272.05K) > 14.7 \text{ kJ/mol} \quad (7.27)$$

### 7.3 Interactive Pruning

As was stated in Chapter 2, one of the techniques for improving the efficiency of a generate and test procedure is to prune partial solutions. Evaluation of partial solutions is often not possible in a design problem. The Interactive Design Procedure is easily cast

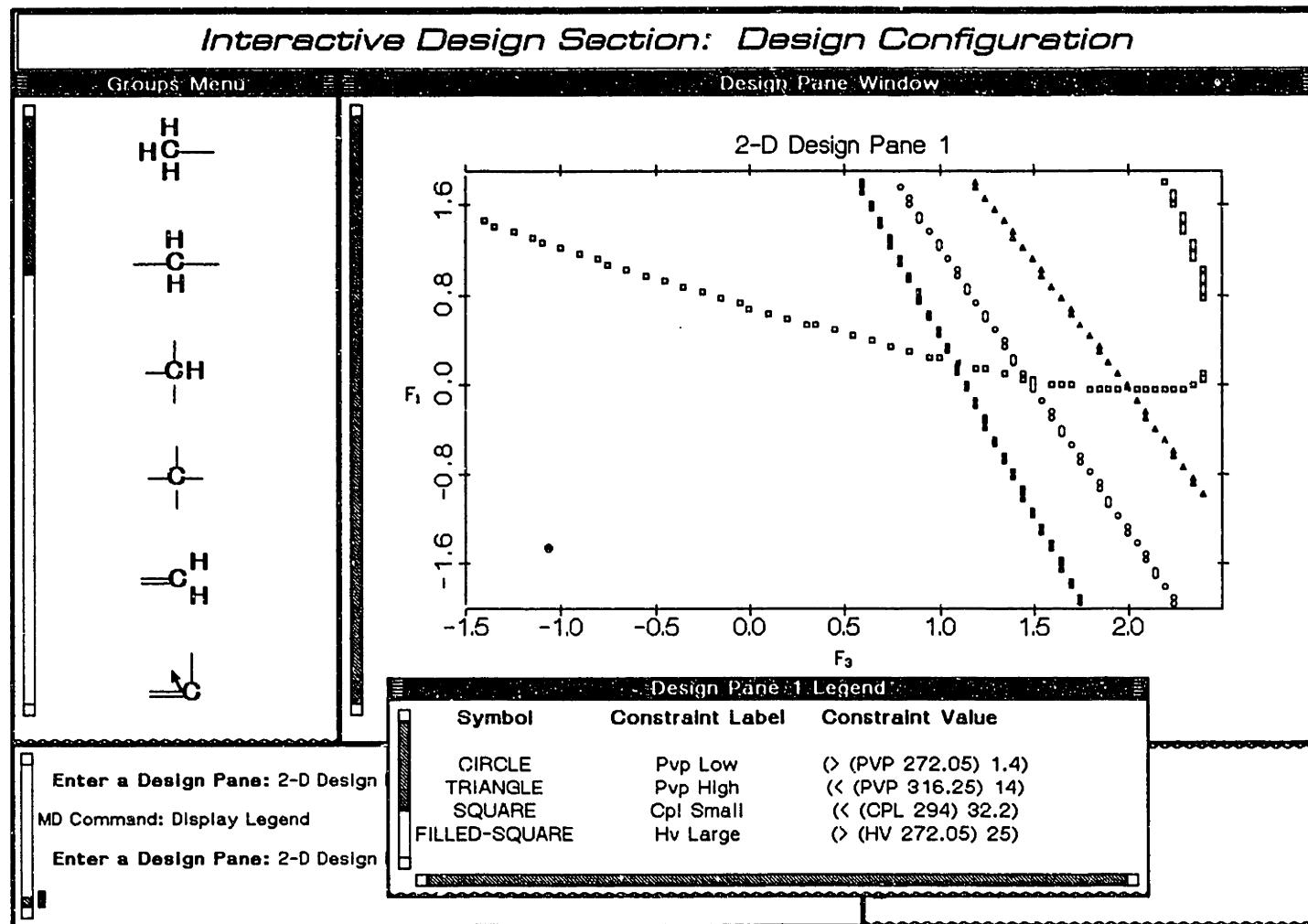


Figure 7.7: Effect of Modifying  $\Delta H_v$  Constraint on the Feasible Region

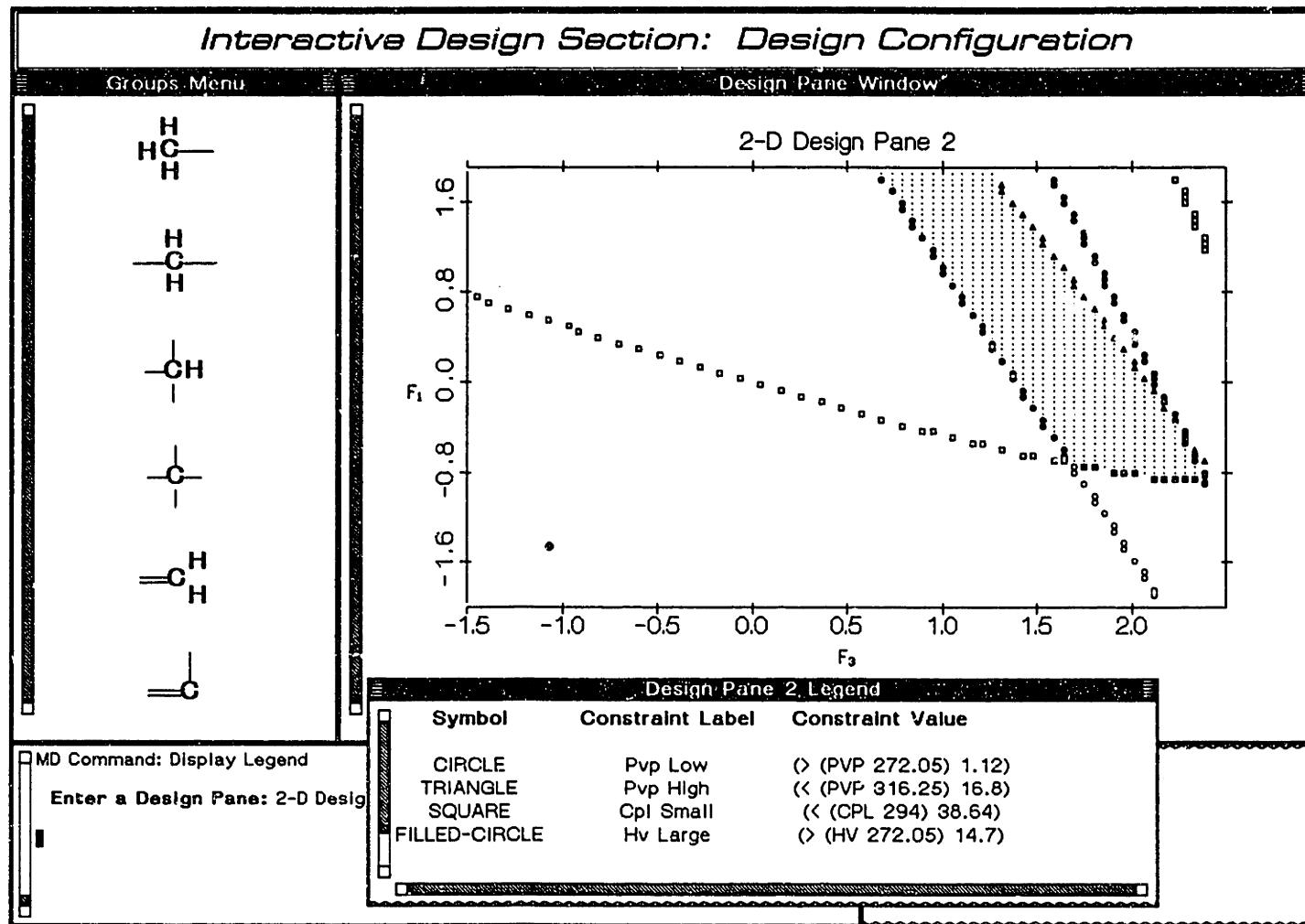


Figure 7.8: Expanded Feasible Region Formed by Relaxing Constraints by 20%

as a generate and test procedure in which the designer uses his or her own knowledge and bias to evaluate and prune partial solutions. The pruning done in the interactive design procedure is heuristic. There is no guarantee that a good solution will not be overlooked.

## 7.4 Cognitive Model of Interactive Design

The interactive design replaces a combinatoric problem of choosing numbers which add up to desired values with a combinatoric problem of choosing vectors which when combined place one within a graphical target. Interestingly the graphical combinatoric problem is easier to solve. To understand why and to identify facilities to assist in the problem's solution I investigated a cognitive model of the interactive design process.

### 7.4.1 Representation

Computer science states that an information-processing system can be understood in terms of its data structures, or representations, and the processes which operate on these representations. The problem of representation is called the problem of *declarative knowledge*. Humans are hypothesized to possess two kinds of declarative knowledge: *propositions*, language-like representations; *images*, perception-like representations[119]. Research on visual imagery centers on the hypothesis that imagery is a special-purpose component of the cognitive architecture containing representations and processes dedicated to processing visual information. This architecture is distinct from the the architecture supporting propositional representations[119].

Kosslyn, et.al.[69] hypothesized that visual imagery is a distinct component of the cognitive architecture with a dedicated representational mechanism. They proposed that visual imagery results from a representation in a short-term *visual buffer*. The buffer is analogous to the *bit-mapped* memory of a computer's monitor screen[119]. Kosslyn[69] hypothesizes that the buffer is a memory structure with intrinsic two-dimensional spatial properties, which is operated on by a set of procedures that can load, refresh, and perform various transformations.

My interactive design procedure represents the design problem graphically. This representation is similar to hypothesized human representation. I believe that without a graphical representation, a molecular designer still constructs a visual image of the numerical group contribution values used in a design. This mentally visualized representation will probably have some deficiencies. Simon[114] has shown that a verbal description of an object is often embellished in memory. Stillings[119] gives the example of imagining a three-inch cube which has two adjacent blue sides. The image in the "mind's eye" will have a particular orientation relative to a point of view. This information was not provided in the original description but is necessary to create a mental image. By explicitly representing group contributions as vectors this need for embellishing is eliminated.

One important characteristic hypothesized for the visual buffer is that the center of the buffer has the highest resolution and the highest level of activation. The center thus receives the bulk of the subject's attention. The result of this behavior is seen in the manner most interactive designs are begun. The distance between the starting point and the target region is reduced by the first several group vectors as much as

possible. Ideally one or two group vectors would bring one near the target region. This allows both the current point of design and the target region both to occupy the central point of focus in the visual buffer.

The representation of group contributions as vectors is beneficial from another point of view. Hubel and Wiesel[60] studied how neurons in a cat's visual cortex respond to different patterns of visual stimulation. Recording electrodes were placed in a cat's visual cortex and simple bars of light were presented to the eye. The bars were moved in various directions, rotated, stretched, and transformed in other ways to find the stimulus that gave rise to the greatest amount of neural activity. Some of the cells fired at the highest frequency when the stimulus was a thin line projected onto a small array of photoreceptors in the cat's retina. Hubel and Wiesel called these cortical cells *line detectors*.

There are neurological disorders of the brain which would make the interactive design method an ineffective representation. *Visual agnosia* is the breakdown of the ability to recognize or identify visually presented objects, although visual function is otherwise intact. Specific forms of visual agnosia have been reported, such as prosopagnosia, the inability to recognize faces. *Spatial neglect* is the tendency to ignore things in a particular region of space regardless of the sensory modality that provides input from that region. Patients with a form of this syndrome called *unilateral spatial neglect* ignore information coming from the right or left of the body midline and may even forget to shave the corresponding side of the face or dress the corresponding side of the body.

### 7.4.2 Focus of Attention

The two concurrent tasks of identifying group vectors which are satisfactory in extent and provide the needed structural features forces a choice in *attention*. Attention can be defined as the selective aspect of perception and response. One of the earliest theories developed to explain selective attention was the *filter* model of Broadbent[14]. The essential notion was that physical features of two stimuli had to pass through a channel of limited capacity.

Since both tasks are essential to the design of structural feasible molecules satisfying physical property constraints it is necessary to provide some facility to allow both tasks to be attended to. We can separate our concurrent task into two compound tasks[10]. The first task is to identify a set of group vectors which satisfy our physical property target. The second task is to check for structural feasibility.

### 7.4.3 Pattern Recognition

A vector has two metrics, angle and magnitude, which can be used for differentiation. Beck[7] investigated how *spatial orientation* directs the effectiveness of similarity in giving rise to grouping or segregation. Objects with the same *orientation* rather than the same shape were seen together. Facilities for clustering group vectors by angle would be beneficial.

Beck's work suggests that the laws of perceptual organization can be reduced basically to organizing principles of brightness and spatial extensity. Beck suggests that processes involving grouping are most sensitive to those properties that are selectively

respondent at an early stage in the visual system. In that context, the processes in grouping are based on spontaneous direct response to relatively simple properties such as *brightness*, *size*, and *line direction*. The basic importance of line orientation is consistent with Gibson[43], who suggested that it is a basic element in the perception of figure. Hubel and Wiesel's[60] works also emphasizes the importance of line recognition.

#### 7.4.4 Zooming

Kosslyn[69] proposed that the human visual imagery system includes many procedures for manipulating the visual buffer. For example, the system can zoom in on a particular region of the buffer, thus expanding that region to fill the entire buffer. The system can also scan the buffer, moving the central region of highest resolution and attention from one area to another.

#### 7.4.5 Design Facilities

Examining the cognitive aspects of the interactive design method leads to the identification of a number of facilities to provide assistance to the designer:

1. **2D Visual Buffer:** The visual buffer hypothesis suggests that design in three dimensions, although possible, should be avoided. The approach I investigated was to project the three dimensional design space onto three two dimensional design spaces.
2. **Group Vector Clustering:** The pattern recognition studies suggest facilities for the clustering of group vectors based upon orientation and size. For group

vectors this translates to clustering by angle and magnitude.

3. **Specifying Groups:** The inability of most humans to dedicate full attention to more than one task at a time suggests the groups available for design should be able to be discriminated. Specifying groups to ensure structural feasibility allows the designer to attend to the single task of identifying a desirable group vector. If no desirable group vectors are found then the focus shifts to identifying a group which will add flexibility to the structure.
4. **Zooming:** Zooming in on the design space would enable the designer to focus his or her visual buffer on the area of the design space involving the current end point and the target region.

#### 7.4.6 Cognitive Sample

To illustrate how the interactive design procedure enhances a human's ability to design molecules because of its emphasis on a cognitive model of human perception I describe a typical thought process followed when performing a design. The thought process I used was my own.<sup>1</sup>

The computer implementation of the interactive design must be described to understand the sample. Figure 7.9 shows the screen of the Symbolics's monitor displaying the implementation of the interactive design procedure. The Design Pane displays a two dimensional  $F_1$ - $F_3$  design space showing the four constraints:

$$P_{vp}(272.05K) > 1.4 \text{ bar} \quad (7.28)$$

---

<sup>1</sup>I make the major (possibly wrong) assumption that my thought process is typical.

$$P_{vp}(316.45K) < 14 \text{ bar} \quad (7.29)$$

$$\Delta H_v(272.05K) > 18.4 \text{ kJ/mol} \quad (7.30)$$

$$C_{p_L}(294.25K) < 32.2 \text{ cal/g-mol}\cdot\text{K} \quad (7.31)$$

The shaded area is the feasible region in which all constraints are satisfied. The Group Menu pane displays the list of groups I have to design with. As the mouse is moved over a displayed group a *temporary* group vector is drawn on the design pane. Figure 7.10 shows a temporary  $-\text{CH}_3$  group.

I begin the design by searching for a group which moves me in the general direction of the target region. This behavior is in accordance with the priority of line orientation noted by Beck[7]. However, I prefer group vectors which bring me closer to the target indicating magnitude is a factor in my decision.

My strategy of beginning first with terminators, then extenders, and then if necessary branchers followed by terminators can be seen as a result of attempting to reduce a concurrent task to a compound task[10]. Instead of having to consider the selection of groups which concurrently satisfy our physical property constraints and form a structurally feasible molecule I simplify the problem. Figure 7.11 shows a facility for restricting the choice of groups. Restrictions can be placed on a number of group characteristics such as a global valence, ring class, and bond type. I restrict the groups to only those having a global valence of one.

Now that I have only terminators to choose from I select my first group from a set of typical groups. For me I generally choose either  $-\text{CH}_3$  or  $-\text{F}$ . This choice comes from my knowledge that it is very likely that either of these groups will occur in a

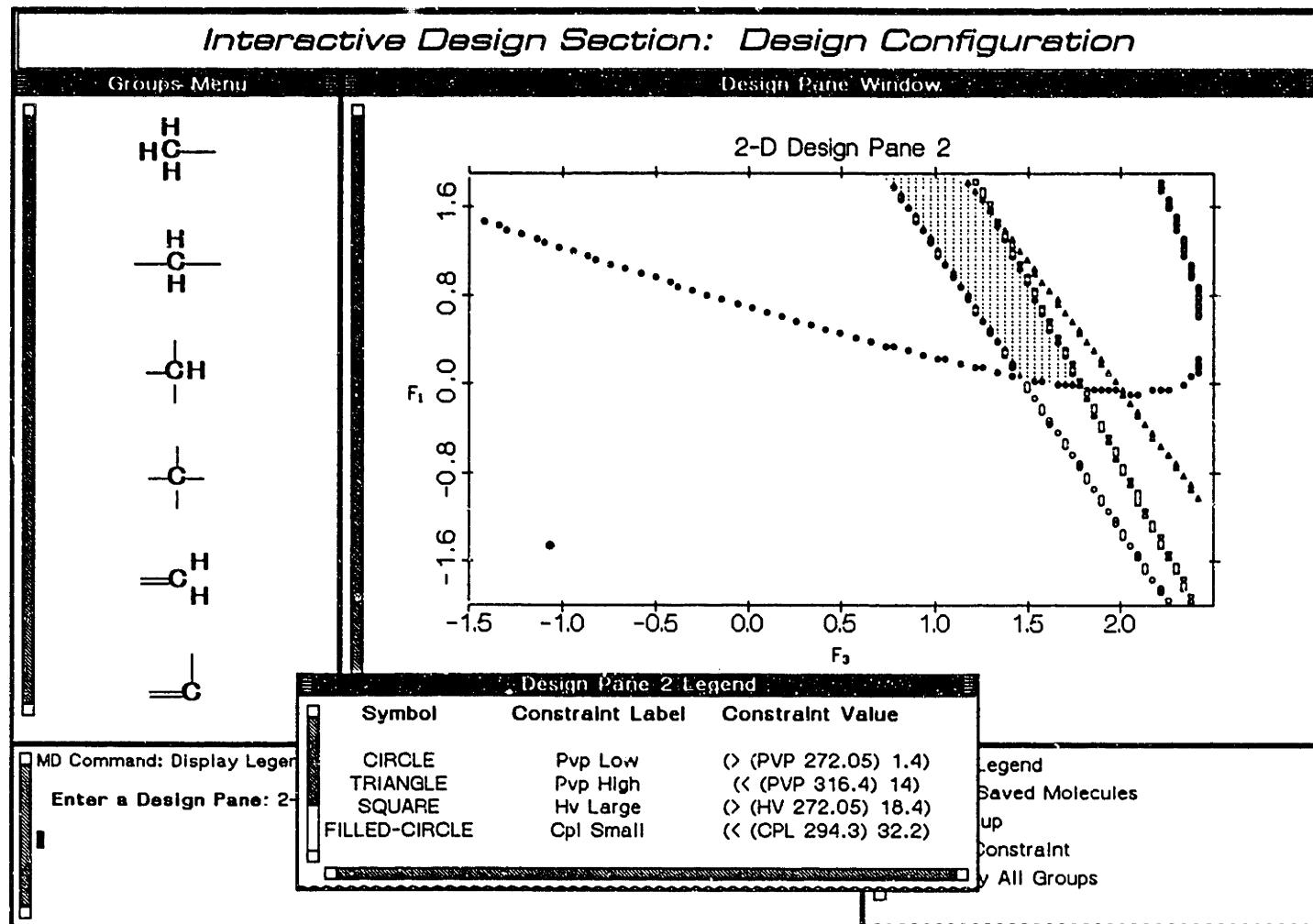


Figure 7.9: Interactive Design Implementation Display

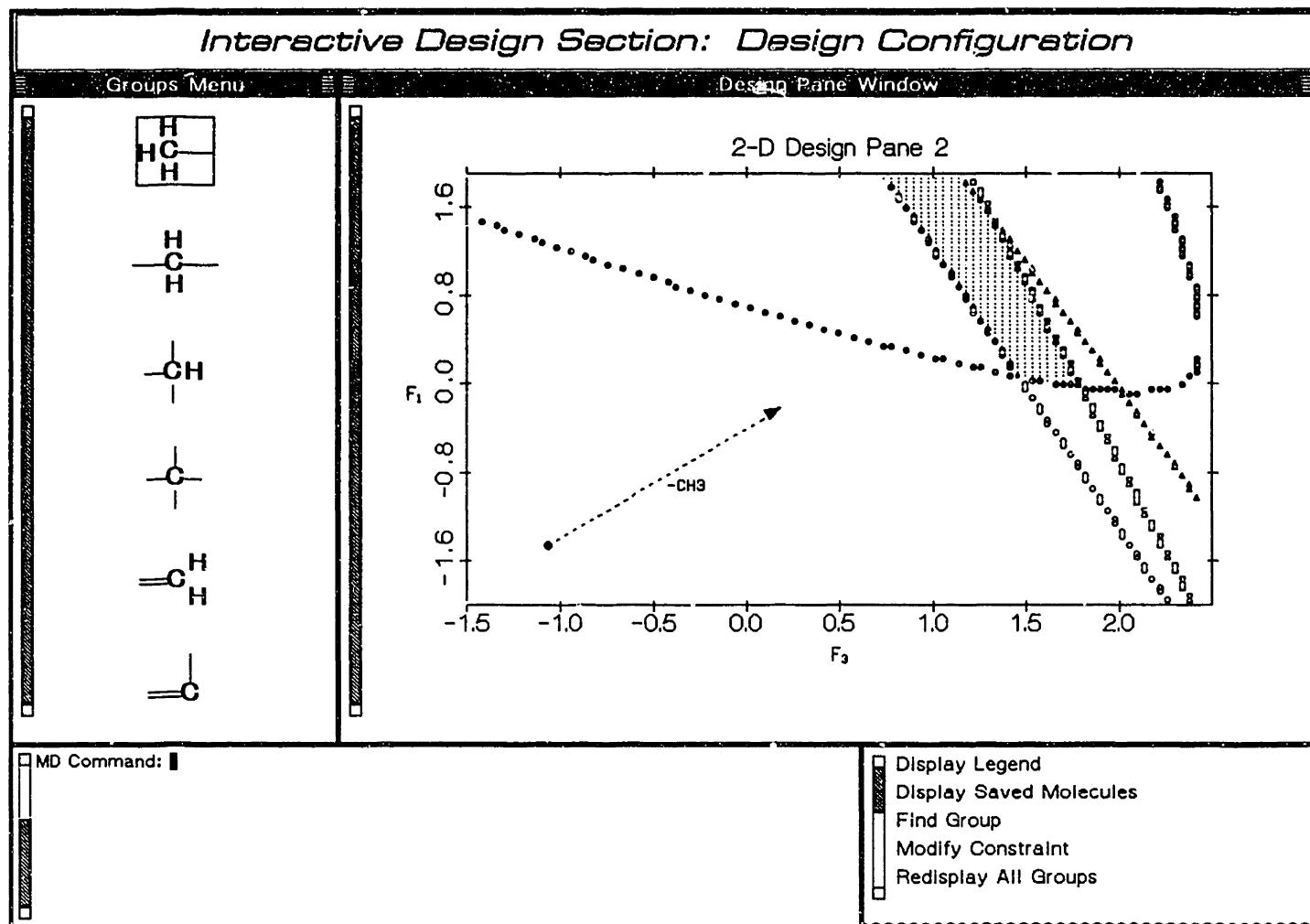


Figure 7.10: Temporary Group Vector

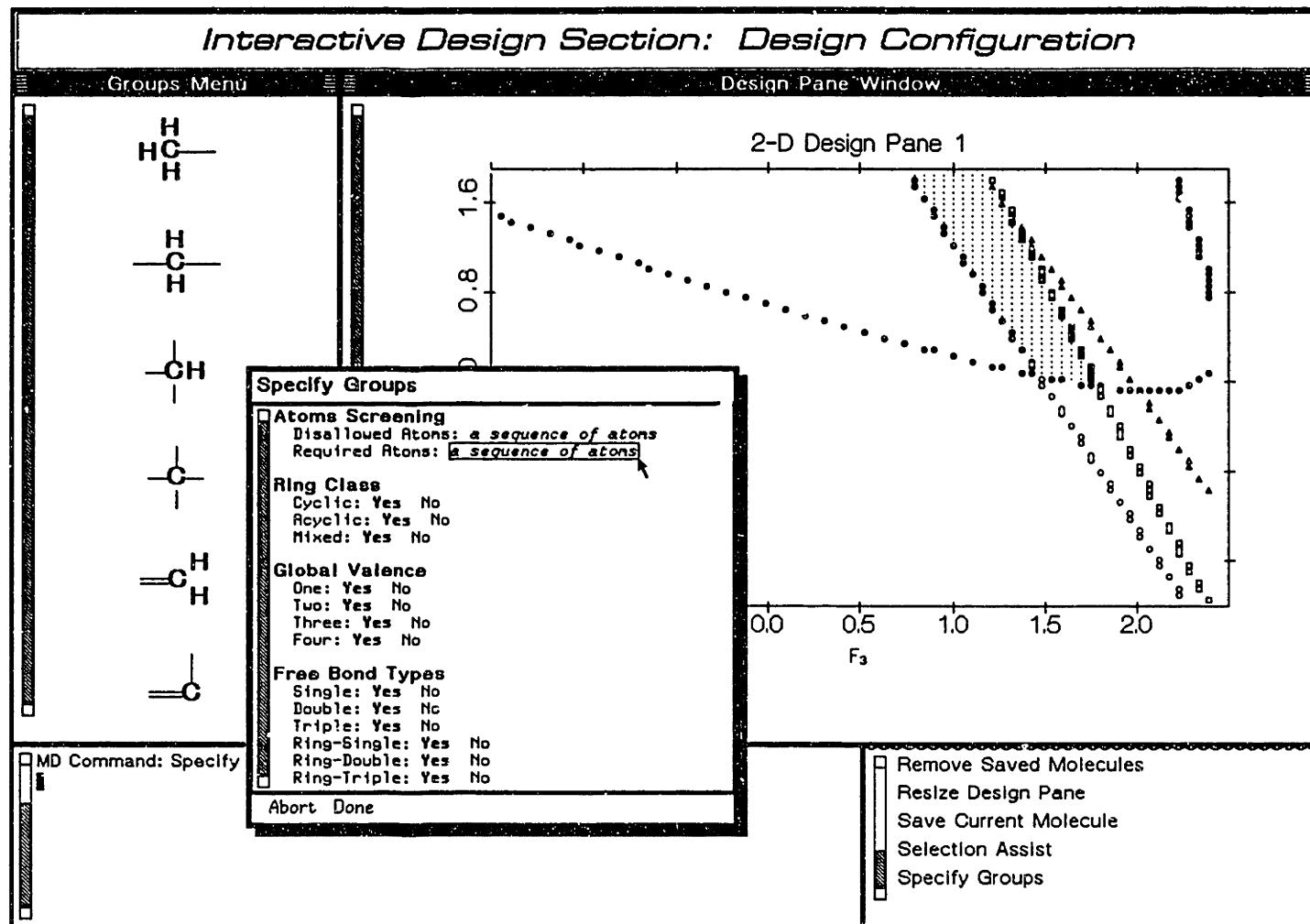


Figure 7.11: Restriction of Displayed Groups

refrigerant. Figure 7.12 shows the effect of choosing the  $-\text{CH}_3$  group. This choice reduces my problem of having to choose two terminators to that of having to choose one. I can now focus my attention on choosing a terminator which brings me close to the target region.

I again place emphasis on the orientation of group vectors. Figure 7.13 shows a facility for restricting the choice of groups. The designer specifies the angle within which group vectors must lie. This allows me to choose a second terminator,  $-\text{Cl}$ . Figure 7.14 shows that these two terminators just satisfy our physical property constraints.

Even though the distribution of group vectors vary widely I feel comfortable being near the target region with a structurally feasible molecule. I believe this satisfaction is due to my ability to focus my attention on a very small region of my visual buffer. To design another molecule I thus attempt to modify the existing  $\text{CH}_3\text{Cl}$  molecule rather than starting another design.

To avoid having to deal with the concurrent constraints of structural feasibility and property satisfaction I restrict the groups to only extenders. This allows me to add groups without affecting their structural feasibility. I again use the facility to specify groups restricting global valence now to two.

Adding a brancher which brings me far from the target region is similar to beginning a design over again. The selection of one or two terminators is again the first step after a brancher has been chosen.

I believe that further use of the interactive design procedure will identify other patterns used by designers in choosing groups. Those I have identified with the aid of an understanding of the cognitive processes involved in group vector selection have led

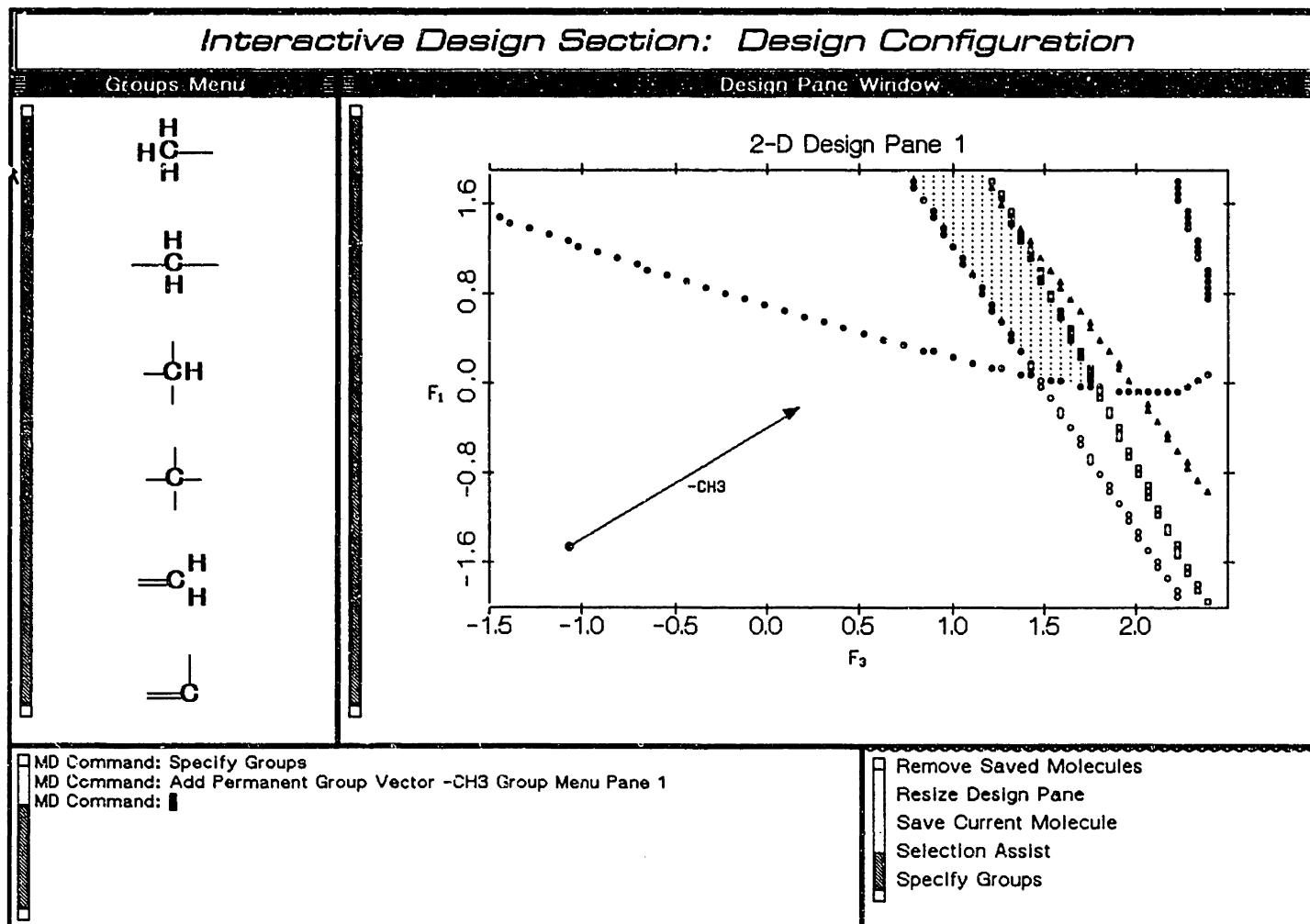


Figure 7.12: Choosing the Methyl Group Vector

### Interactive Design Section: Design Configuration

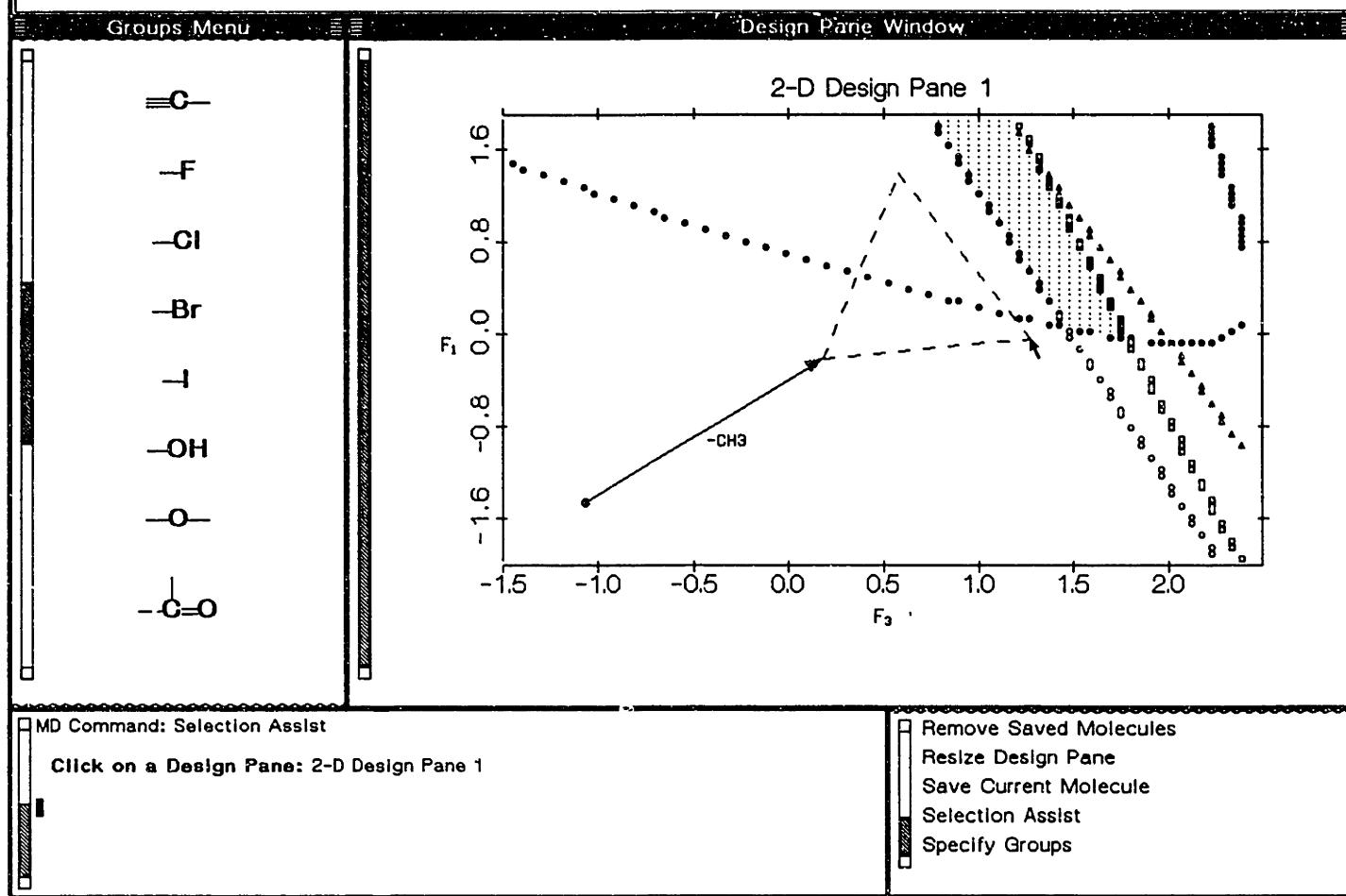


Figure 7.13: Angle Restriction on Group Vectors

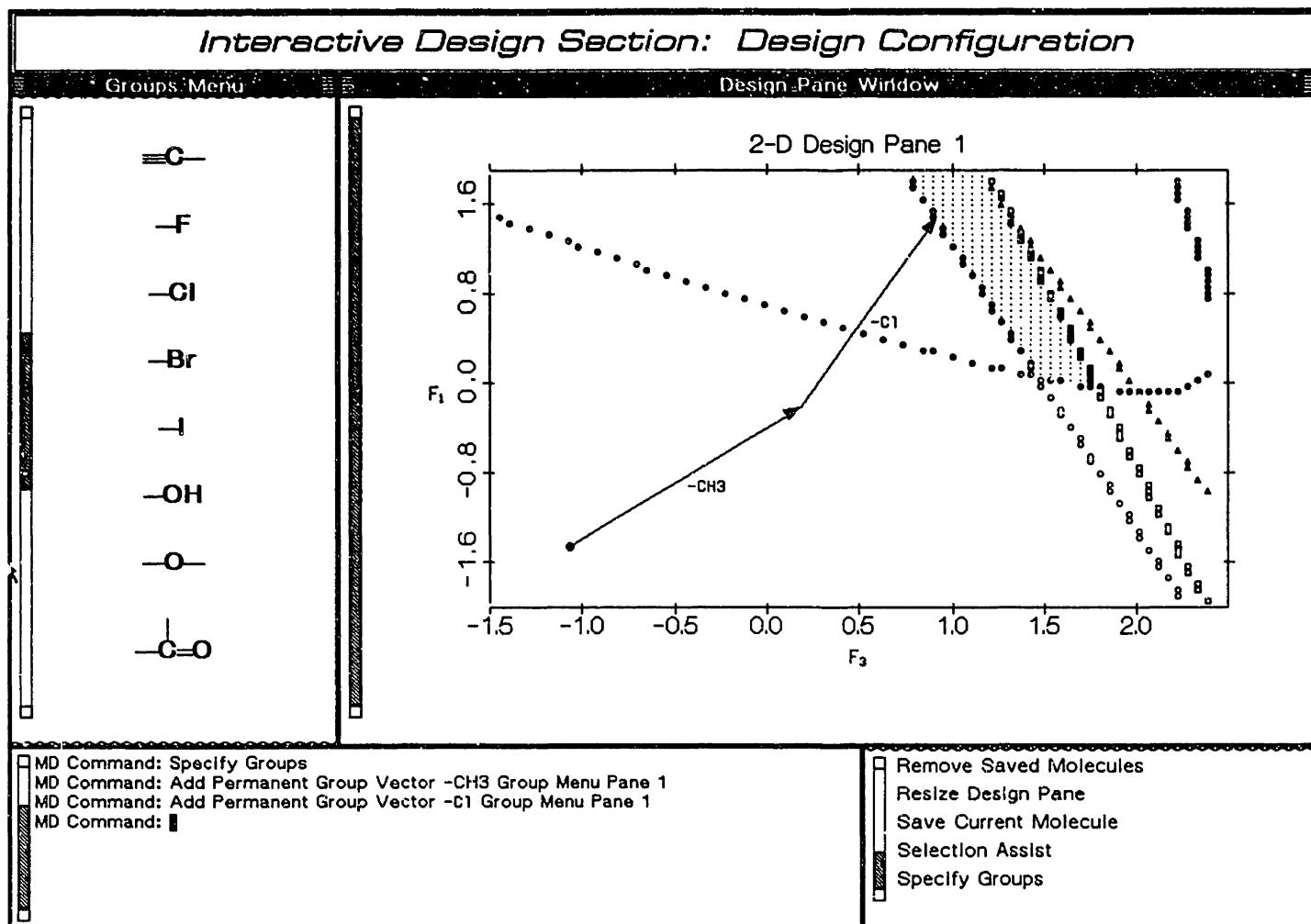


Figure 7.14: Choosing a Second Terminator

to the implementation of facilities which have greatly enhanced the effectiveness of the interactive design procedure.

# Chapter 8

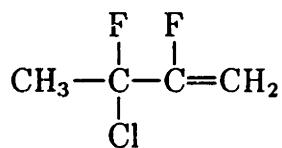
## Enumeration

The result of either the interactive or automatic design procedures is a collection of groups. To form complete molecules it is necessary to connect these groups together. At times more than one way of connecting the groups is possible. All possible structures must then be enumerated.

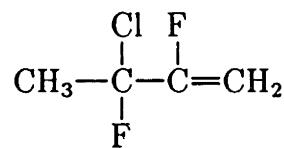
Much work was done in the field of molecular structure enumeration[47]. Generation of all possible isomeric structures given a molecular formula is an important problem in chemistry. I developed a procedure which took advantage of the group representation I had used in my design procedures. The procedure is based on the generate and test paradigm. Since the possible ways of connecting a reasonable number of groups is enormous, it is necessary to improve the efficiency of the search. The constraints available are suitable for pruning partial solutions.

In this chapter I describe the procedure developed for enumeration of complete molecular structures given a set of groups. Because of problems verifying the results the procedure was never fully tested. I describe the procedure and the verification

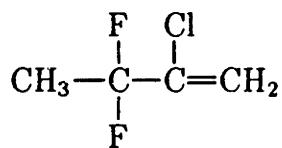
Table 8.1: Four Enumerated Molecules



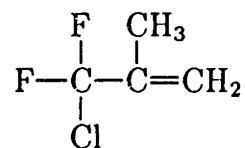
Enumerated Molecule A



Enumerated Molecule B



Enumerated Molecule C

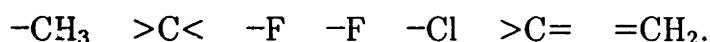


Enumerated Molecule D

difficulty.

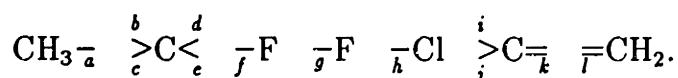
## 8.1 The Problem

The result of either the interactive or automatic design procedure is a collection of groups. An example result might be the following collection of groups:



These groups can be combined to form four different molecules. These four molecules are shown in Table 8.1.

The procedure begins by labeling each of the free bonds in our collection of groups. The labeling is done using letters. The labeled groups follow:



All the bonds, denoted by letters, are collected into a single list:

( a b c d e f g h i j k l ).

Bonds are associated into *proto-molecules*. I denote a proto-molecule by a list of bond pairs. Thus if bond *a* was to be connected with bond *f*, the proto-molecule showing this association would be:

$( ( a f ) )$

Proto-molecules are then pruned using a set of constraints. These constraints test whether the bonds associated are allowed to be connected. There are five constraints:

1. Compatible Types.
2. All Bonds Connected.
3. No Intra-Group Bonding.
4. One Inter-Group Bond.
5. No Duplicates.

The procedure for generating all possible bond associations and the manner in which the constraints are used to prune proto-molecules is described using the example of enumerating the possible molecules formed from the set of labeled groups shown above.

## 8.2 Combinatorics

The number of ways in which two bonds can be chosen from a set of *n* bonds is given by:

$$\frac{n!}{2(n-2)!}. \quad (8.1)$$

Once the first two bonds are selected, there are:

$$\frac{(n-2)!}{2(n-4)!} \quad (8.2)$$

Table 8.2: Combinatorics of Bond Association

Number of Bonds	Number of Associations
4	3
6	15
8	105
10	945
12	10,395
14	135,135
16	2,027,025
18	34,459,425
20	654,729,075
22	13,749,310,575

ways of choosing the next bond pair. Thus, the total number of bond associations is given by:

$$\frac{n!}{2(n-2)!} \frac{(n-2)!}{2(n-4)!} \frac{(n-4)!}{2(n-6)!} \dots \quad (8.3)$$

Canceling the denominator and numerator of consecutive terms:

$$\frac{n!}{2^{n/2}}. \quad (8.4)$$

However, Equation 8.4 considers the order of the bond associations to be important. Pairing bonds  $(a\ b)$  and then  $(c\ d)$  is different than pairing  $(c\ d)$  and then  $(a\ b)$ . In molecule enumeration the order is not important. Equation 8.4 is thus reduced by the number of permutations of the  $n/2$  bond associations:

$$\frac{n!}{2^{n/2} (n/2)!}. \quad (8.5)$$

Table 8.2 shows values for the number of bond associations for a given number of bonds.

## 8.3 Enumeration Procedure

All the bonds of the labeled groups are first collected into a list:

$(a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l).$

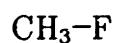
The procedure begins by forming all possible proto-molecules with the first bond in the list as one of the bonds. Eleven possible proto-molecules are formed:

$((a\ b))\ ((a\ c))\ ((a\ d))\ ((a\ e))$   
 $((a\ f))\ ((a\ g))\ ((a\ h))\ ((a\ i))$   
 $((a\ j))\ ((a\ k))\ ((a\ l))$

Each of these proto-molecules is tested using the enumeration constraints.

**Compatible Types Enumeration Constraint:** Bonds must be of a compatible type to be connected. Single bonds can only connect to single bonds, double with double, etc. Bond  $a$  represents a single bond. Bonds  $k$  and  $l$  represent double bonds. These bonds can not be connected. Thus, proto-molecules  $((a\ k))$  and  $((a\ l))$  are pruned away.

**All Bonds Connected Enumeration Constraint:** All the bonds must be used in the enumeration procedure. The proto-molecule  $((a\ f))$  represents connecting group  $\text{CH}_3\bar{a}$  with group  $\bar{f}\text{F}$ . This forms the complete molecule:



leaving no place for attaching any of the remaining bonds. This constraint prunes proto-molecules  $((a\ f))$ ,  $((a\ g))$ , and  $((a\ h))$  away.

**No Duplicates Enumeration Constraint:** Identification of duplicate proto-molecules is a difficult task. The four proto-molecules:

$((a\ b))$     $((a\ c))$     $((a\ d))$     $((a\ e))$

are all considered duplicates. This is because with three bonds free on the  $>C<$  group all four bonds are considered identical. However, if bonds  $b$  and  $c$  were connected then bonds  $d$  and  $e$  may or may not be identical. This depends if the attachments to bonds  $b$  and  $c$  are identical and leads to the question of whether or not chirality is considered in identifying duplicates. I did not account for chirality. Thus, the remaining two bonds,  $d$  and  $e$ , would be considered identical.

Each group is examined for identical bonds. All free bonds of one type connected to the same atom are identical. Thus, the bonds  $i$  and  $j$  are identical in the  $\overset{i}{>}C=\overset{j}{k}$  group. However, the bonds  $m$  and  $n$  are not identical in the  $\overset{m}{-}COO\overset{n}{-}$  group because they are not connected to the same atom.

The procedure for finding duplicates is to identify all identical bonds on a group and substitute one of the bonds for all the others in all proto-molecules. Then identify all duplicate groups and substitute one bond for all the others in all proto-molecules. Proto-molecules whose bonds lists match are duplicates. When duplicates are identified one is chosen and the remaining are pruned away. Thus in the set of duplicates:

$((a\ b))$     $((a\ c))$     $((a\ d))$     $((a\ e))$ ,

proto-molecule  $((a\ b))$  is kept and the others pruned away. In the set of duplicates:

$((a\ i))$     $((a\ j))$ ,

proto-molecule  $((a\ i))$  is kept and proto-molecule  $((a\ j))$  is pruned away.

Two proto-molecules remain:

$$((a\ b)) \quad ((a\ i)).$$

For each proto-molecule the bonds still needing to be connected are collected into a list. For proto-molecule  $((a\ b))$  the list of remaining bonds is:

$$(c\ d\ e\ f\ g\ h\ i\ j\ k\ l).$$

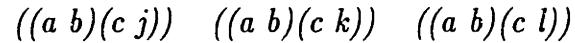
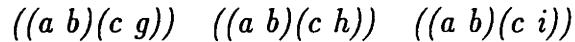
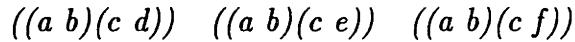
The association procedure is repeated with the bond associations checked by the above three enumeration constraints.

There are nine possible associations of bond  $c$  with another bond. Each of these associations is combined with the proto-molecule  $((a\ b))$  to form a new set of proto-molecules. These are:

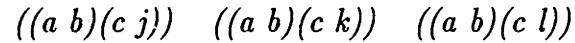
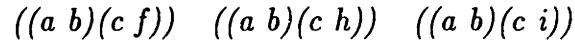
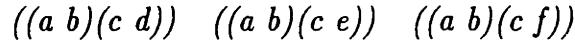
$$\begin{aligned} & ((a\ b)(c\ d)) \quad ((a\ b)(c\ e)) \quad ((a\ b)(c\ f)) \\ & ((a\ b)(c\ g)) \quad ((a\ b)(c\ h)) \quad ((a\ b)(c\ i)) \\ & ((a\ b)(c\ j)) \quad ((a\ b)(c\ k)) \quad ((a\ b)(c\ l)) \end{aligned}$$

The Compatible Types Enumeration Constraint prunes proto-molecules  $((a\ b)(c\ k))$  and  $((a\ b)(c\ l))$ . The All Bonds Connected Enumeration Constraint does not prune any proto-molecules. The No Duplicates Enumeration Constraint prunes proto-molecules  $((a\ b)(c\ e))$ ,  $((a\ b)(c\ g))$ , and  $((a\ b)(c\ j))$ .

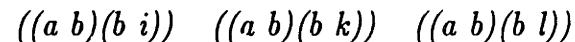
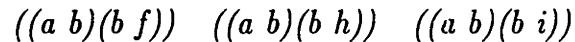
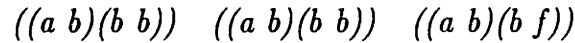
The application of the No Duplicates Enumeration Constraint is demonstrated here. The initial set of proto-molecules is:



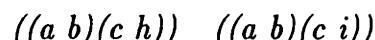
The single duplicate group is  $-F$ . From the two bonds  $f$  and  $g$  one is chosen. I chose bond  $f$ . Bond  $g$  is replaced with bond  $f$  in all the proto-molecules yielding:



Groups  $\overset{b}{>}C\overset{d}{<}e$  and  $\overset{i}{>}C\overset{e}{=}\overset{j}{k}$  have identical bonds. From the bonds  $b$ ,  $c$ ,  $d$ , and  $e$  one is chosen. I chose bond  $b$ . Bond  $b$  replaces bonds  $b$ ,  $c$ ,  $d$ , and  $e$  in all the proto-molecules. From the bonds  $i$  and  $j$  I chose bond  $i$ . Bond  $j$  is replaced with bond  $i$ . The resulting proto-molecules are:



We see there are three duplicates in the set of proto-molecules. Comparing these with the original set we find the proto-molecules to prune. The remaining proto-molecules are:



The No Intra-Group Bonding constraint is now applied.

**No Intra-Group Bonding Enumeration Constraint:** Two bonds on the same group can not be connected. If all the bonds of a group are replaced with one bond intra-group bonding is identified by bond associations which consist of the same bond. For example, replacing bonds  $c$ ,  $d$ , and  $e$  with bond  $b$  results in proto-molecule  $((a\ b)(c\ d))$  becoming  $((a\ b)(b\ b))$ . The bond association  $(b\ b)$  signals intra-group bonding. Applying this constraint prunes proto-molecule  $((a\ b)(c\ d))$ .

Performing the same procedure on the proto-molecule  $((a\ i))$  we obtain the set of possible proto-molecules:

$$\begin{array}{lll} ((a\ b)(c\ f)) & ((a\ b)(c\ h)) & ((a\ b)(c\ i)) \\ ((a\ i)(b\ f)) & ((a\ i)(b\ h)) & ((a\ i)(b\ j)) \end{array}$$

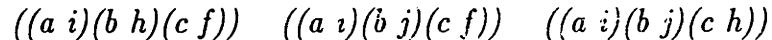
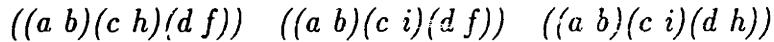
The third set of bond associations is now formed. After pruning by the three enumeration constraints:

1. Compatible Types.
2. All Bonds Connected.
3. No Intra-Group Bonding.

the following set of proto-molecules remain:

$$\begin{array}{lll} ((a\ b)(c\ f)(d\ g)) & ((a\ b)(c\ f)(d\ h)) & ((a\ b)(c\ f)(d\ i)) \\ ((a\ b)(c\ h)(d\ f)) & ((a\ b)(c\ h)(d\ i)) & \\ ((a\ b)(c\ i)(d\ f)) & ((a\ b)(c\ i)(d\ h)) & ((a\ b)(c\ i)(d\ j)) \\ ((a\ i)(b\ f)(c\ g)) & ((a\ i)(b\ f)(c\ h)) & ((a\ i)(b\ f)(c\ j)) \\ ((a\ i)(b\ h)(c\ f)) & ((a\ i)(b\ h)(c\ j)) & \\ ((a\ i)(b\ j)(c\ f)) & ((a\ i)(b\ j)(c\ h)) & \end{array}$$

Applying the No Duplicates enumeration constraint prunes six proto-molecules:

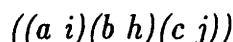
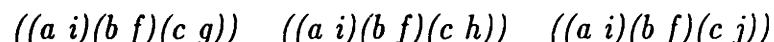
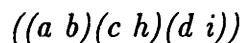


The final constraint to be applied is the One Inter-Group Bond enumeration constraint.

**One Inter-Group Bond Enumeration Constraint:** Only one bond is allowed to connect any two groups. Proto-molecule  $((a\ b)(c\ i)(d\ j))$  has the groups  $\overset{b}{>}C\overset{d}{<}_{e}$  and  $\overset{i}{>}C\overset{i}{=}_{k}$  joined by two bonds. This is not allowed.

To identify proto-molecules which have more than one inter-group bond we replace all bonds of a group with one bond. Bonds  $b$ ,  $c$ ,  $d$ , and  $e$  are all on group  $\overset{b}{>}C\overset{d}{<}_{e}$  and are replaced with bond  $b$ . Bonds  $i$ ,  $j$ , and  $k$  are all on group  $\overset{i}{>}C\overset{i}{=}_{k}$  and are replaced with bond  $i$ . After substitution proto-molecule  $((a\ b)(c\ i)(d\ j))$  becomes  $((a\ b)(b\ i)(b\ i))$ . Any proto-molecule which has the same bond association occurring more than once has more than one inter-group bond and is pruned away.

The eight remaining proto-molecules are:



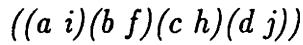
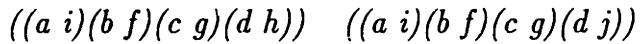
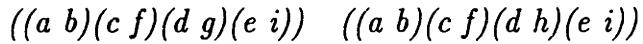
The bond association process is repeated three more times. After the generation of proto-molecules the enumeration constraints are applied. The three sets of resulting proto-molecules are shown in Table 8.3. The final set of proto-molecules corresponds to the four molecules shown in Table 8.1. The proto-molecule  $((a\ b)(c\ f)(d\ h)(e\ i)(g$

---

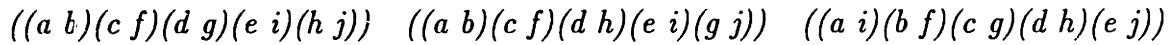
Table 8.3: Proto-Molecules

---

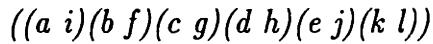
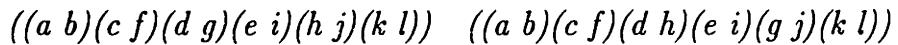
Four Bond Associations



Five Bond Associations



Six Bond Associations



---

$j)(k\ l))$  corresponds to both Enumerated Molecules A and B since the enumeration algorithm does not account for chirality.

## 8.4 Implementation Difficulties

The major difficulty encountered in the implementation of the enumeration procedure was the display of the final complete molecules. Especially during development, display of the results was essential to verifying and extending the algorithm. The automatic

display of molecular structures is a difficult problem. In Chapter 16 I discuss the previous work in molecular structure display, and recommendations I have for future research.

# Chapter 9

## Molecule Screening

The molecule enumeration step generates complete molecules. After enumeration it is possible to effectively use chemical constraints. Many chemical constraints need to examine sections of molecules larger than the group scale. The chemical constraints I have considered are disallowed substructures. This is the approach taken by DENDRAL[76]. The major task in applying these kinds of chemical constraints is identification of molecular substructures.

In this chapter I describe the procedure I developed for the identification of molecular substructures. I did not complete the development of the procedure due to implementational difficulties.

### 9.1 The Problem

*Substructure search* is the problem of determining whether or not a particular molecular substructure is present in a molecule. The most common application of substructure

Table 9.1: Disallowed Substructures

$>\text{C}=\text{C}-\text{NH}-$	$>\text{C}=\text{C}-\text{OH}$	$=\text{C}=\text{C}-\text{NH}-$
$=\text{C}=\text{C}-\text{OH}$	$-\text{C}\equiv\text{C}-\text{NH}-$	$-\text{C}\equiv\text{C}-\text{OH}$
$-\text{N}=\text{N}-\text{N}<$	$-\text{N}=\text{N}-\text{N}=$	$-\text{N}=\text{N}-\text{O}-$
$-\text{N}=\text{NH}$	$-\text{CH}-\text{N}=\text{O}$	$-\text{N}=\overset{ }{\text{C}}-\text{OH}$
$-\text{O}-\text{O}-$	$>\text{N}-\overset{ }{\text{N}}-\text{N}<$	$=\text{N}-\overset{ }{\text{N}}-\text{N}<$
$=\text{N}-\overset{ }{\text{N}}-\text{N}=$	$>\text{N}-\text{O}-\text{N}<$	$=\text{N}-\text{O}-\text{N}<$
$=\text{N}-\text{O}-\text{N}=$	$-\text{O}-\overset{ }{\text{N}}-\text{N}<$	$-\text{O}-\overset{ }{\text{N}}-\text{N}=$
$-\text{O}-\text{S}-$	$-\text{S}-\text{S}-$	$-\text{NH}-\overset{ }{\text{C}}-\text{NH}_2$
$-\text{NH}-\overset{  }{\text{C}}-\text{NH}_2$	$-\text{O}-\text{COOH}$	$>\text{N}-\text{COOH}$
$=\text{N}-\text{COOH}$		

search routines is in systems for searching computer files of structures for those that contain a specified substructure[47]. There are two main methods for searching for specified substructures[47]: 1) the set reduction technique proposed by Sussenguth[121]; 2) the node-by-node technique proposed by Ray and Kirsch[99].

## 9.2 Disallowed Substructures

DENDRAL possessed a list of substructures which were disallowed in proposed molecular structures. These forbidden substructures are shown in Table 9.1. The system examines each of the candidate molecules to see if it contains a disallowed substructure. If the molecule does contain the substructure it is removed from further consideration.

The major need is to provide a means for matching molecular substructures. A number of methods have been published. I examined these and decided to investigate a possible new method which could take advantage of some of the capabilities of LISP.

### 9.3 Substructure Representation

To apply chemical constraints in the form of disallowed substructures it is necessary to develop a procedure for identifying substructures. The first step in developing the procedure is to develop a representation for molecules and substructures for use in matching. I represent molecular structures by a list of bonds.

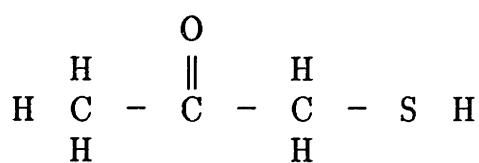
Most of the algorithms used in molecular structure manipulation concentrated on atoms using a connection table as the means of representation. I concentrate on bonds. Bonds have a consistency which atoms do not. All bonds connect two and only two atoms.

Each element of the bond list is a five member list of the form:

*(bond-type atom-1 atom-1-identifier atom-2 atom-2-identifier)*.

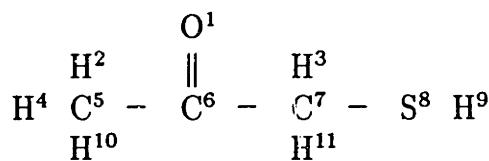
The meaning of each of these terms is explained with the aid of an example.

Given a molecule such as:

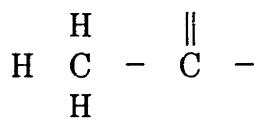


the first step is to assign an atom-identifier to each of the atoms. For molecules I simply use the positive integers as atom identifiers. It should be emphasized that the integers

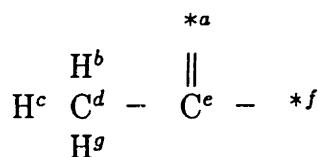
are used for the purpose of distinguishing atoms of the same type and can be arbitrarily assigned. The molecule with atoms identified is:



The substructure being searched for is also assigned atom-identifiers. For substructures I use letters. The substructure is:



The substructure with the atoms identified is:



Since substructures represent only partial molecular structures there will be unconnected bonds. For consistency of representation I connect these bonds to “wildcard” atoms. These represent any kind of atom. I denote a wildcard atom by an \*.

Bond lists are formed for both the substructure and molecule. These are shown in Table 9.2.

The present form of the representation does not account for chirality. The two substructures shown in Figure 9.3 have the same bond list. Thus, even though the substructures are different in three dimensional space, they are represented identically.

At this time I have not examined whether or not this representation can represent chirality. I believe it can. Chirality is simply an ordering of the bonds in three dimensional space. One way of representing this ordering is to order the bonds in the

Table 9.2: Example Bond Lists

Bond Lists	
Substructure	Molecule
(double * a C e)	(double O 1 C 6)
(single H b C d)	(single H 2 C 5)
(single H c C d)	(single H 3 C 7)
(single C d H g)	(single H 4 C 5)
(single C d C e)	(single C 5 H 10)
(single C e * f)	(single C 5 C 6)
	(single C 6 C 7)
	(single C 7 H 11)
	(single C 7 S 8)
	(single S 8 H 9)

Table 9.3: Two Substructures in Fisher Projections

$\text{F}^1$	$\text{Br}^5$
$\text{Cl}^2 \text{ C}^3 - *^4$	$\text{Cl}^2 \text{ C}^3 - *^4$
$\text{Br}^5$	$\text{F}^1$
(single F 1 C 3) (single Cl 2 C 3) (single C 3 * 4) (single Br 5 C 3)	

Table 9.4: Molecule – Substructure Matching Pair

$  \begin{array}{ccccccccc}  & & \text{O}^1 & & & & & & \\  \text{H}^2 & \text{C}^5 & - & \parallel & \text{H}^3 & & & \text{H}^b & \text{C}^d & - & \parallel & \text{C}^e & - & *f \\  \text{H}^4 & \text{C}^6 & - & & \text{C}^7 & - & \text{S}^8 & \text{H}^c & \text{H}^g & & & & & \\  & \text{H}^{10} & & & \text{H}^{11} & & & & & & & & & & \\  & & \text{Molecule F} & & & & & & & & & & & & \text{Substructure A}  \end{array}  $	$*a$ $*b$ $*c$ $*d$ $*e$ $*f$ $*g$
--	--

bond list. This necessitates being able to distinguish between bonds of similar type. The representation would then be extended to the number of atoms and bonds in a structure. Each element of the bond list would then be of the form:

*(bond-type bond-identifier atom-1 atom-1-identifier atom-2 atom-2-identifier).*

## 9.4 Substructure Identification Procedure

The identification procedure proceeds in a hierarchical fashion in a manner similar to Sussenguth's[121]. There are four steps to the matching procedure:

1. Match atom and bond types.
2. Match atoms and bonds.
3. Match individual atom–bond connections.
4. Match all atom–bond connections.

Each step of the matching procedure requires more information than the previous one. Proceeding in a hierarchical manner should thus minimize the effort needed in finding a match. The four steps of the procedure are demonstrated using the molecule and substructure of Figure 9.4.

#### 9.4.1 Stage 1: Matching Atom and Bond Types

For a substructure to be present in a molecule the molecule must contain the same types of atoms and bonds as the substructure. Substructure A contains two types of atoms:

C H

and two types of bonds:

The “wildcard” atoms  ${}^a$  and  ${}^f$  are not considered as atoms in this listing.

Molecule F contains four types of atoms:

C H O S

and two types of bonds:

Since both the atom set and bond set of Substructure A are subsets of the atom set and bond set of Molecule F they are considered to match at this stage of the procedure.

#### 9.4.2 Stage 2: Matching Atoms and Bonds

Once the types of atoms and bonds match, the next step is to ensure there are a sufficient number of them. The matched atoms and bonds of Stage 1 are counted in both Substructure A and Molecule F. Table 9.5 shows the results of this counting.

Table 9.5: Atom and Bond Counts

	Substructure A	Molecule F
Atom Type	Atom Count	Atom Count
C	2	3
H	3	6
O	0	1
S	0	1
Bond Type	Bond Count	Bond Count
Single	5	9
Double	1	1

For each atom and bond type the number of occurrences in the substructure must be less than or equal to the number of occurrences in the molecule. The occurrences of all the atom and bond types in Substructure A are less than in Molecule F. The substructure and molecule thus match at this stage.

#### 9.4.3 Stage 3: Match Individual Atom–Bond Connections

The next stage of the matching procedure checks that individual atom–bond connections match in both the substructure and molecule. In essence this check is done between the bond lists of the substructure and molecule. Ignoring atom identifiers, each element of the bond list of the substructure should match each element of the bond list of the molecule. This matching procedure is simply done by “subtracting” each element of the bond list of the substructure from the bond list of the molecule. If a subtraction can not be performed, there is no matching bond element in the molecule’s bond list, then the substructure is not included in the molecule.

For a bond element to match at this stage only the bond types and atoms must match. For the bond lists in Table 9.2 we have the following subtraction:

(double * a C e)	→	(double O 1 C 6)
(single H b C d)	→	(single H 2 C 5)
(single H c C d)	→	(single H 3 C 7)
(single C d H g)	→	(single H 4 C 5)
		(single C 5 H 10)
(single C d C e)	→	(single C 5 C 6)
(single C e * f)	→	(single C 6 C 7)
		(single C 7 H 11)
		(single C 7 S 8)
		(single S 8 H 9)

Since all the bond elements of Substructure A's bond list could be subtracted from the bond list of Molecule F, Substructure A is considered to be included in Molecule F at this stage.

In the above subtraction we matched the bond element:

(single H c C d)

of the substructure with the bond element:

(single H 3 C 7)

of the molecule. Referring to Table 9.4 we see that these two bonds do not actually correspond to each other. However, since we are ignoring the atom identifiers at this stage this match is allowed. The final stage of the matching procedure is when the atom identifiers are considered.

#### 9.4.4 Stage 4: Match All Atom–Bond Connections

In this final stage all atom–bond connections are matched. In Stage 3 we did not account for the atom identifiers. This allowed us to match bonds which were inconsistent.

A carbon atom connected to three hydrogen atoms could be matched with any three carbon-hydrogen bond elements. In this stage all the bond matchings must form a consistent set.

The procedure at this stage begins by simply checking all possible assignments to each of the atom-identifiers in the substructure until a match is found. We begin with one of the bond elements in Substructure A:

(single H b C d).

Each of the bond elements of Molecule F is checked for a match with the bond type and the atoms types. Bond element:

(single H 2 C 5)

of Molecule F matches the bond element of Substructure A. The atom identifiers of Substructure A are now assigned the values from Molecule F:

b → 2

d → 5

These assignments are propagated through the remaining bond elements of Substructure A's bond list. The updated bond list of Substructure A is:

(double \* a C e)  
(single H 2 C 5)  
(single H c C 5)  
(single C 5 H g)  
(single C 5 C e)  
(single C e \* f)

A new bond element is now chosen to be resolved with the a bond element in Molecule F.

The bond element chosen is:

(double \* a C e).

This can be resolved with only one bond element of Molecule F's bond list:

(double O 1 C 6).

The assignments made to the atom identifiers are:

a → 1

e → 6.

Propagating these assignments Substructure A's bond list becomes:

(double \* 1 C 6)  
(single H 2 C 5)  
(single H c C 5)  
(single C 5 H g)  
(single C 5 C 6)  
(single C 6 \* f)

The two assignments have completed the resolution for another bond element:

(single C 5 C 6).

This bond element of Substructure A is checked against the bond elements of Molecule F. Molecule F contains a matching bond element and the procedure continues. Another bond element is chosen from Substructure A's bond list and is resolved against a bond element of Molecule F. When a match for all the atom identifiers in a substructure is found the substructure is included in the molecule. If all possible assignments were tried without success then the substructure is not included.

The detailed procedure for searching through all the possibilities has not been constructed. Many questions need to be answered. However, I believe the bond list representation serves as a good basis for beginning this research.

The idea of this procedure comes from the process of resolution in predicate calculus I believe investigation into the implementation of resolution theorem provers will provide insight into improving the performance of this algorithm.

## 9.5 Group Formation

Besides its use in the application of chemical constraints, a procedure for the matching of substructures would be useful in the development of group contribution estimation techniques. One of the major difficulties in the formation of the multiple regressions used in development of such techniques is the identification of the occurrences of groups within the data set of molecules.

The typical procedure used is to pick a set of groups and then enter the occurrences of these groups in the molecules by hand. This process has the limitation that it is very difficult to vary the groups used in the regression analysis. A change in the groups would mean the occurrence data would have to be recoded.

# Chapter 10

## Molecule Evaluation

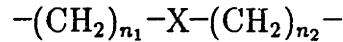
It is sometimes necessary to modify estimation techniques for use in the generate and test procedures. Often this modification is done to remove steps in the estimation techniques which require knowledge of global molecular structure. Using groups as my design basis I only know local structure during the design. However, once the candidate molecules are enumerated and screened, global molecular structure is known and more accurate estimation techniques can be used to further prune the candidates.

Van Krevelen[127] developed an group contribution estimation technique for  $T_g$ . The contributions for several groups contained correction terms accounting for their “environment”. For example, the  $\text{--COO--}$ group’s contribution was:

$$8000 + 12000 I_x$$

$I_x$  is an *interaction factor* representing the “linear concentration” of polar groups within a flexible chain of methylene beads.  $I_x$  is defined as the number of main chain atoms in the polar group,  $X$ , divided by the number of chain atoms of this group plus those

of the directly connected methylene chains. For the configuration:



in which the characteristic group **X** contains  $n_x$  chain atoms, the formula of  $I_x$  is:

$$I_x = \frac{n_x}{n_x + n_1 + n_2}. \quad (10.1)$$

The knowledge required to compute  $I_x$  is not available at the time of design. The corrections for polar groups was not made. After the candidate molecules have been designed and enumerated there is global knowledge about the molecular structure to use the unmodified estimation technique.

In this final step of the methodology the most rigorous estimation techniques are used to evaluate and possibly prune the candidate molecules. Rigorous estimation techniques include molecular modeling.

# Chapter 11

## Refrigerant Design

Many current refrigerants deplete the Earth's protective ozone layer. Automotive air conditioners are a major source of refrigerant emissions. In this case study I examined the design of replacement refrigerants for automotive air conditioners. Thermodynamic analysis of the refrigeration process identified important physical property constraints. Using these constraints I performed both interactive and automatic designs. Molecules designed by both methods are presented.

### 11.1 Refrigeration

Refrigeration is the process of transferring heat from a low temperature to a high temperature at the expense of work. The principle methods of refrigeration available are:

1. Vapor recompression
2. Vapor absorption
3. Air cycle

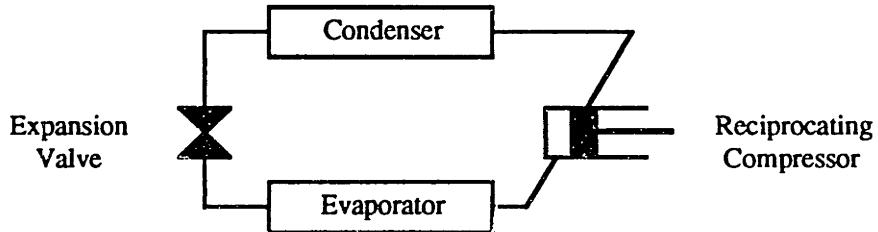


Figure 11.1: Basic Refrigeration System

#### 4. Thermo-electric

Automotive air-conditioners use the vapor recompression principle.

Figure 11.1 shows a schematic of a basic vapor recompression refrigeration system. The thermodynamic processes which occur within the system are best depicted on a Pressure-Enthalpy diagram. A hypothetical P-H diagram is shown in Figure 11.2.

In the evaporator the refrigerant at low temperature contacts the process stream needing cooling. The refrigerant is saturated liquid at this point and is denoted by state **F** on Figure 11.2. As the liquid absorbs heat it evaporates changing to a saturated vapor at state **A**. This vapor exits the evaporator and enters the compressor. Theoretically this compression is performed adiabatically resulting in a hot vapor at state **B**. This vapor is cooled isobarically to state **C** and then condensed to a saturated liquid at state **D**. The heat rejected by the refrigerant is absorbed by some sink usually cooling water or air. The cooled, high pressure, saturated liquid is now flashed isenthalpically through an expansion valve resulting in a cold mixture of vapor and liquid at state **E**. The refrigerant separates into a liquid phase at state **F** and a vapor phase at state **A**. The liquid at state **F** begins the cycle again.

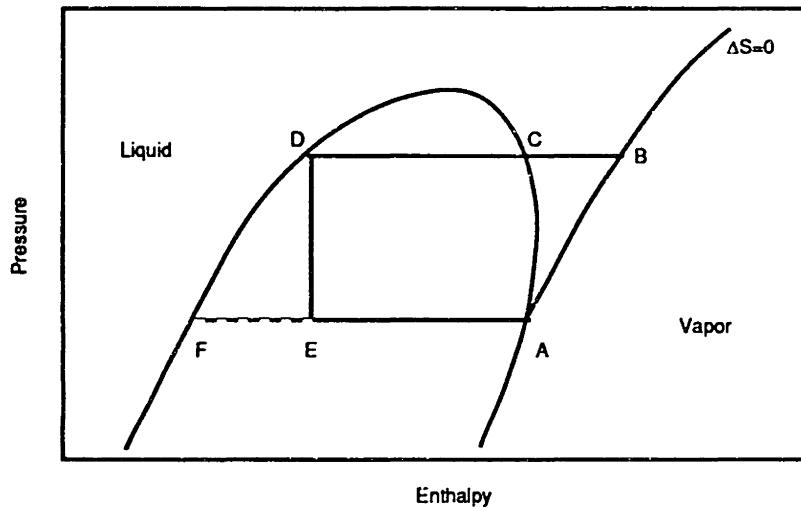


Figure 11.2: Hypothetical Refrigeration Cycle

## 11.2 Current Refrigerants

Refrigerants fall into three general categories[113]:

1. Saturated and unsaturated aliphatic hydrocarbons, e.g., propane and propylene.
2. Aliphatic halogenated hydrocarbons, e.g., dichlorodifluoromethane.
3. Inorganic gases, e.g., air,  $\text{CO}_2$ ,  $\text{SO}_2$ ,  $\text{NH}_3$ , and  $\text{Cl}_2$ .

The major class of compounds which have found use as refrigerants are the halogenated hydrocarbons specifically chlorofluorocarbons(CFC's).

For the saturated and unsaturated aliphatic hydrocarbons and aliphatic halogenated hydrocarbons, a numerical coding system describes the refrigerant's molecular structure. The general formula is:

$$\text{R} - W X Y Z$$

in which:

$W$  = the number of double bonds.

$X$  = (the number of carbon atoms) - 1.

$Y$  = (the number of hydrogen atoms) + 1.

$Z$  = the number of fluorine atoms.

Thus dichlorodifluoromethane is denoted by R-12. Inorganic refrigerants are assigned three-digit numbers, the first of which is 7 with the following two numbers being the compound's molecular weight. Thus,  $\text{NH}_3$ , which has a molecular weight of 17, is denoted by R-717.

Table 11.1 lists a number of fluids having properties which render them suitable for use as refrigerants.

### 11.3 Problems with Chlorofluorocarbons

Chlorofluorocarbons(CFC's) deplete the Earth's ozone layer. CFC's decompose releasing chlorine atoms. In the stratosphere chlorine atoms react with ozone to form  $\text{ClO}$ . Chlorine monoxide reacts with atomic oxygen to regenerate the chlorine atom, thereby completing a catalytic cycle which destroys ozone[134]. This catalytic degradation of ozone is shown in the following reactions[116]:

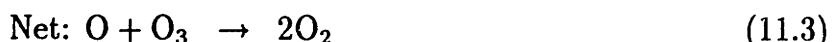


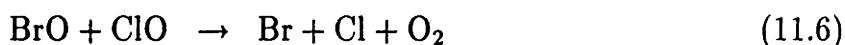
Table 11.1: Current Refrigerants

ASRE Standard Refrigerant Designation	Chemical Name	Chemical Formula	Molecular Weight	Boiling Point °C
10	Carbon Tetrachloride	CCl <sub>4</sub>	153.8	76.8
11	Trichloromonofluoromethane	CCl <sub>3</sub> F	137.4	26.0
12	Dichlorodifluoromethane	CCl <sub>2</sub> F <sub>2</sub>	120.9	29.8
13	Monochlorotrifluoromethane	CClF <sub>3</sub>	104.5	-81.4
13B1	Monobromotrifluoromethane	CBrF <sub>3</sub>	148.9	-57.8
14	Carbontetrafluoride	CF <sub>4</sub>	88.0	-128.0
20	Chloroform	CHCl <sub>3</sub>	119.4	61.0
21	Dichloromonofluoromethane	CHCl <sub>2</sub> F	102.9	8.9
22	Monochlorodifluoromethane	CHClF <sub>2</sub>	86.5	-40.8
23	Trifluoromethane	CHF <sub>3</sub>	70.0	-84.4
30	Methylene Chloride	CH <sub>2</sub> Cl <sub>2</sub>	84.9	40.7
31	Monochloromonofluoromethane	CH <sub>2</sub> ClF	68.5	8.9
32	Methylene Fluoride	CH <sub>2</sub> F <sub>2</sub>	52.0	-51.9
40	Methyl Chloride	CH <sub>3</sub> Cl	50.5	-23.8
41	Methyl Fluoride	CH <sub>3</sub> F	34.0	-78.3
50	Methane	CH <sub>4</sub>	16.0	-161.7
110	Hexachloroethane	CCl <sub>3</sub> CCl <sub>3</sub>	236.8	185.0
111	Pentachloromonofluoroethane	CCl <sub>3</sub> CCl <sub>2</sub> F	220.3	137.0
112	Tetrachlorodifluoroethane	CCl <sub>2</sub> FCCl <sub>2</sub> F	203.8	92.8
112a	Tetrachlorodifluoroethane	CCl <sub>3</sub> CClF <sub>2</sub>	203.8	91.0
113	Trichlorotrifluoroethane	CCl <sub>2</sub> FCClF <sub>2</sub>	187.4	47.6
113a	Trichlorotrifluoroethane	CCl <sub>3</sub> CF <sub>3</sub>	187.4	45.7
114	Dichlorotetrafluoroethane	CClF <sub>2</sub> CClF <sub>2</sub>	170.9	3.6
114a	Dichlorotetrafluoroethane	CCl <sub>2</sub> FCF <sub>3</sub>	170.9	3.6
114B2	Dibromotetrafluoroethane	CBrF <sub>2</sub> CBrF <sub>2</sub>	259.9	47.6
115	Monochloropentafluoroethane	CClF <sub>2</sub> CF <sub>3</sub>	154.5	-38.7
116	Hexafluoroethane	CF <sub>3</sub> CF <sub>3</sub>	138.0	-78.2
120	Pentachloroethane	CHCl <sub>2</sub> CCl <sub>3</sub>	202.3	162.2
123	Dichlorotrifluoroethane	CHCl <sub>2</sub> CF <sub>3</sub>	153.0	28.7
124	Monochlorotetrafluoroethane	CHClFCF <sub>3</sub>	136.5	-12.0
124a	Monochlorotetrafluoroethane	CHF <sub>2</sub> CClF <sub>2</sub>	136.5	-10.0
125	Pentafluoroethane	CHF <sub>2</sub> CF <sub>3</sub>	120.0	-48.3
133a	Monochlorotrifluoroethane	CH <sub>2</sub> ClCF <sub>3</sub>	118.5	6.1
140a	Trichloroethane	CH <sub>3</sub> CCl <sub>3</sub>	133.4	73.9
142b	Monochlorodifluoroethane	CH <sub>3</sub> CClF <sub>2</sub>	100.5	-11.0
143a	Trifluoroethane	CH <sub>3</sub> CF <sub>3</sub>	84.0	-47.5

Table 11.1 Continued: Current Refrigerants

ASRE Standard Refrigerant Designation	Chemical Name	Chemical Formula	Molecular Weight	Boiling Point °C
150a	Dichloroethane	CH <sub>3</sub> CHCl <sub>2</sub>	98.9	60.0
152a	Difluoroethane	CH <sub>3</sub> CHF <sub>2</sub>	66.0	-24.7
160	Ethyl Chloride	CH <sub>3</sub> CH <sub>2</sub> Cl	64.5	12.2
170	Ethane	CH <sub>3</sub> CH <sub>3</sub>	30.0	-88.6
218	Octafluoropropane	CF <sub>3</sub> CF <sub>2</sub> CF <sub>3</sub>	188.0	-38.0
290	Propane	CH <sub>3</sub> CH <sub>2</sub> CH <sub>3</sub>	44.0	-42.3
717	Ammonia	NH <sub>3</sub>	17.0	-33.3
718	Water	H <sub>2</sub> O	18.0	100.0
729	Air		29.0	-194.4
744	Carbon Dioxide	CO <sub>2</sub>	44.0	-78.3 subl.
744A	Nitrous Oxide	N <sub>2</sub> O	44.0	-88.3
764	Sulfur Dioxide	SO <sub>2</sub>	64.0	-0.0

Bromine undergoes similar reaction in addition to a combined reaction with chlorine[86]:



Although many compounds possess chlorine as an atomic constituent, CFC's are seen as a major problem because they are stable enough to reach the upper atmosphere. Most other compounds decompose much lower in the atmosphere and thus pose no direct threat to the ozone layer. The group contribution of the fluorine atom toward  $\Delta G_{f,298}^\circ$  is  $-247.19 \text{ kJ/mol}$ [62]. Thus although chlorine is the main component in the reactive degradation of ozone, fluorine is also responsible because it greatly stabilizes

the compound.

## 11.4 Problem: Automotive Air Conditioning

The specific refrigerant design problem I examined is the design of a refrigerant for use in an automotive air conditioning cycle. The compressors used in automotive air conditioning units are of the reciprocating type. The design temperatures for which I design the refrigerant are a high temperature of 110°F and a low temperature of 30°F[71].

## 11.5 Freon 12

The most commonly used refrigerant for automobile air conditioning is dichlorodifluoromethane or Freon-12. The important properties of Freon-12 at the conditions of interest are listed in Table 11.2.

## 11.6 Problem Formulation

Identifying the target is the first step in my methodology. The following constraints specify the desired physical properties of an automotive air conditioning refrigerant:

- $P_{vp}(T = -1.1^\circ\text{C}) > 1.4 \text{ bar}$

The lowest pressure in the cycle should be greater than atmospheric[28]. This reduces the possibility of air and moisture leaking into the system. Douglas[29] recommends a safety factor of 5 psig.

Table 11.2: Physical Properties of Freon-12

Property	Value	Units
$M_w$	120.914	
$T_f$	115.4	K
$T_b$	243.4	K
$T_c$	385.0	K
$P_c$	40.7	atm
$\omega$	0.176	
$\Delta H_{vb}$	4772.	cal/g-mol
$\Delta H_{f,298}^\circ$	-115.0	kcal/g-mol
$\Delta G_{f,298}^\circ$	-105.7	kcal/g-mol

Property	Saturated	Saturated	Units
	Vapor	Liquid	
Temperature = 30°F			
Vapor Pressure	2.97	2.97	bar
Volume	6.94	8.63e-2	$\text{m}^3/\text{kg}\cdot\text{mol}$
Enthalpy	2.262e+7	4.235e+6	$\text{J}/\text{kg}\cdot\text{mol}$
Viscosity	9.54e-8	2.35e-6	$\text{kg}\cdot\text{mol}/\text{m}\cdot\text{sec}$
Thermal Conductivity	2.46e-3	2.49e-2	$\text{J}/\text{m}\cdot\text{K}\cdot\text{sec}$
Heat Capacity	7.77e+4	1.12e+5	$\text{J}/\text{kg}\cdot\text{mol}\cdot\text{K}$
Temperature = 110°F			
Vapor Pressure	10.42	10.42	bar
Volume	2.02	9.76e-2	$\text{m}^3/\text{kg}\cdot\text{mol}$
Enthalpy	2.47e+7	9.43e+6	$\text{J}/\text{kg}\cdot\text{mol}$
Viscosity	1.13e-7	1.57e-6	$\text{kg}\cdot\text{mol}/\text{m}\cdot\text{sec}$
Thermal Conductivity	3.31e-3	1.93e-2	$\text{J}/\text{m}\cdot\text{K}\cdot\text{sec}$
Heat Capacity	9.72e+4	1.24e+5	$\text{J}/\text{kg}\cdot\text{mol}\cdot\text{K}$

- $P_{vp}(T = 43.3^\circ\text{C}) < 14 \text{ bar}$

A high system pressure increases the size, weight, and cost of equipment[28]. A pressure ratio of 10 is considered to be the maximum for a refrigeration cycle.[94].

- $\Delta H_v(T = -1.1^\circ\text{C}) > 18.4 \text{ kJ/g-mol}$

A large enthalpy of vaporization reduces the amount of refrigerant needed. The value for freon-12's enthalpy of vaporization at  $-1.1^\circ\text{C}$  is  $18.4 \text{ kJ/g-mol}$ [3].

- $C_{pL}(T = 21.1^\circ\text{C}) < 32.2 \text{ cal/g-mol}\cdot\text{K}$

It is desirable to have a low liquid heat capacity[28]. A low liquid heat capacity reduces the amount of refrigerant which flashes upon passage through the expansion valve. The heat capacity constraint is evaluated at an average temperature.

Freon-12's liquid heat capacity at  $21.1^\circ\text{C}$  is  $32.2 \text{ cal/g-mol}\cdot\text{K}$ [3].

With these constraints elucidated the problem formulation step of the design procedure is now complete. The next step is to identify the estimation techniques which are used to determine the physical property values. This identification process is done in the next step: target transformation.

## 11.7 Target Transformation

Estimation procedures are developed for each of the physical properties used in our target constraints:

$$P_{vp} \quad H_v \quad C_{pL}.$$

I design refrigerants using both the interactive and automatic procedures. Two transformed targets are thus needed, one for each design procedure.

### 11.7.1 Transformation for Interactive Design

Estimation procedures were developed for interactive design in two dimensions. These estimation procedures are presented here.<sup>1</sup> <sup>2</sup>

**Vapor Pressure Estimation Procedure:** The vapor pressure is used in two constraints:

$$P_{vp}(T = -1.1^\circ\text{C}) > 1.4 \text{ bar} \quad (11.8)$$

$$P_{vp}(T = 43.3^\circ\text{C}) < 14 \text{ bar.} \quad (11.9)$$

The estimation procedure for the vapor pressure used in the interactive design follows:

- 1)  $P_{vp} = P_{vp}(T_b, T_c, P_c)$  by Riedel-Plank-Miller EOT
- 2)  $T_b = T_b(F_1, F_2, F_3)$  by  $T_b$  Factor EOT
- 2)  $T_c = T_c(F_1, F_2, F_3)$  by  $T_c$  Factor EOT
- 2)  $P_c = P_c(F_1, F_2, F_3)$  by  $P_c$  Factor EOT
- 3)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
- 3)  $F_2 = 0$  by  $F_2$  assumption
- 3)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

The resulting fundamental properties of the estimation procedure are  $F_1$  and  $F_3$ .

---

<sup>1</sup> Appendix A describes the estimation techniques used.

<sup>2</sup> Appendix B evaluates the accuracy of estimation procedures.

**Enthalpy of Vaporization:** The enthalpy of vaporization is used in one constraint:

$$\Delta H_v(T = -1.1^\circ\text{C}) > 18.4 \text{ kJ/g-mol.} \quad (11.10)$$

The estimation procedure for the enthalpy of vaporization used in the interactive design follows:

- 1)  $\Delta H_v = \Delta H_v(\Delta H_{vb}, T_b, T_c)$  by Watson EOT
- 2)  $\Delta H_{vb} = \Delta H_{vb}(F_1, F_2, F_3)$  by  $\Delta H_{vb}$  Factor EOT
- 2)  $T_b = T_b(F_1, F_2, F_3)$  by  $T_b$  Factor EOT
- 2)  $T_c = T_c(F_1, F_2, F_3)$  by  $T_c$  Factor EOT
  - 3)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
  - 3)  $F_2 = 0$  by  $F_2$  assumption
  - 3)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

The resulting fundamental properties of the estimation procedure are  $F_1$  and  $F_3$ .

**Liquid Heat Capacity:** The liquid heat capacity is used in one constraint:

$$C_{pL}(T = 21.1^\circ\text{C}) < 32.2 \text{ cal/g-mol}\cdot\text{K.} \quad (11.11)$$

The estimation procedure for the liquid heat capacity used in the interactive design follows:

- 1)  $C_{pL} = C_{pL}(\omega, C_p^\circ, T_c)$  by Rowlinson EOT
- 2)  $\omega = \omega(T_{br}, P_c)$  by Lee-Kesler EOT
- 2)  $C_p^\circ \approx C_p^\circ(298\text{K})$  by  $C_p$  approximation
- 3)  $T_{br} = T_{br}(T_c, T_b)$  by definition
  - 4)  $C_p^\circ(298\text{K}) = C_p^\circ(F_1, F_2, F_3)$  by  $C_p^\circ$  Factor EOT
  - 4)  $T_c = T_c(F_1, F_2, F_3)$  by  $T_c$  Factor EOT

Table 11.3: Consistent Groups for Refrigerant Design

$-\text{CH}_3$	$-\text{CH}_2-$	$>\text{CH}-$	$>\text{C}<$	$=\text{CH}_2$
$=\text{CH}_2$	$=\text{CH}<$	$=\text{C}=$	$\equiv\text{CH}$	$\equiv\text{C}-$
$-\text{F}$	$-\text{Cl}$	$-\text{Br}$	$-\text{I}$	$-\text{OH}$
$-\text{O}-$	$>\text{C}=\text{O}$	$\text{O}=\text{CH}-$	$-\text{COOH}$	$=\text{O}$
$-\text{NH}_2$	$>\text{NH}$	$>\text{N}-$	$-\text{CN}$	$-\text{NO}_2$
$-\text{SH}$	$-\text{S}-$			

- 4)  $T_b = T_b(F_1, F_2, F_3)$  by  $T_b$  Factor EOT
- 5)  $F_1 = F_1(\text{groups})$  by Joback  $F_1$  GCT
- 5)  $F_2 = 0$  by  $F_2$  assumption
- 5)  $F_3 = F_3(\text{groups})$  by Joback  $F_3$  GCT

$C_p^\circ$  is approximated by the heat capacity at 298K,  $C_p^\circ(298\text{K})$ . This is to relate the heat capacity to the three factors. The two resulting fundamental properties are  $F_1$  and  $F_3$ .

### 11.7.2 Resulting Properties and Groups

The resulting fundamental properties of the estimation procedure are  $F_1$  and  $F_3$ . Table 11.3 displays the consistent groups for the group contribution techniques for  $F_1$  and  $F_3$ .

### 11.7.3 Transformation for Automatic Design

The number of fundamental properties is not important when performing an automatic design. Estimation procedures are developed which yield the greatest accuracy. I developed the following estimation procedures for automatic design.

**Vapor Pressure Estimation Procedure:** The vapor pressure is used in two constraints:

$$P_{vp}(T = -1.1^\circ\text{C}) > 1.4 \text{ bar} \quad (11.12)$$

$$P_{vp}(T = 43.3^\circ\text{C}) < 14 \text{ bar.} \quad (11.13)$$

The estimation procedure for the vapor pressure used in the automatic design follows:

- 1)  $P_{vp} = P_{vp}(T_b, T_{br}, P_c)$  by Riedel-Plank-Miller EOT
- 2)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(\text{groups})$  by Joback  $T_{br}$  GCT
- 2)  $P_c = P_c(\text{groups})$  by Joback  $P_c$  GCT

The resulting fundamental properties of the estimation procedure are  $T_b$ ,  $T_{br}$ , and  $P_c$ .

**Enthalpy of Vaporization:** The enthalpy of vaporization is used in one constraint:

$$\Delta H_v(T = -1.1^\circ\text{C}) > 18.4 \text{ kJ/g-mol.} \quad (11.14)$$

The estimation procedure for the enthalpy of vaporization used in the automatic design follows:

- 1)  $\Delta H_v = \Delta H_v(\Delta H_{vb}, T_b, T_{br})$  by Watson EOT
- 2)  $\Delta H_{vb} = \Delta H_{vb}(\text{groups})$  by Joback  $\Delta H_{vb}$  GCT
- 2)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(\text{groups})$  by Joback  $T_{br}$  GCT

The resulting fundamental properties of the estimation procedure are  $T_b$ ,  $T_{br}$ , and  $\Delta H_{vb}$ .

**Liquid Heat Capacity:** The liquid heat capacity is used in one constraint:

$$C_{pL}(T = 21.1^\circ\text{C}) < 32.2 \text{ cal/g-mol}\cdot\text{K}. \quad (11.15)$$

The estimation procedure for the liquid heat capacity used in the automatic design follows:

- 1)  $C_{pL} = C_{pL}(\omega, C_p^\circ, T_c)$  by Rowlinson EOT
- 2)  $\omega = \omega(T_{br}, P_c)$  by Lee-Kesler EOT
- 2)  $T_c = T_c(T_b, T_{br})$  by definition
- 2)  $C_p^\circ = C_p^\circ(C_{p,a}^\circ, C_{p,b}^\circ, C_{p,c}^\circ, C_{p,d}^\circ)$  by  $C_p^\circ$  cubic fit
  - 3)  $T_{br} = T_{br}(\text{groups})$  by Joback  $T_{br}$  GCT
  - 3)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
  - 3)  $C_{p,a}^\circ = C_{p,a}^\circ(\text{groups})$  by Joback  $C_{p,a}^\circ$  GCT
  - 3)  $C_{p,b}^\circ = C_{p,b}^\circ(\text{groups})$  by Joback  $C_{p,b}^\circ$  GCT
  - 3)  $C_{p,c}^\circ = C_{p,c}^\circ(\text{groups})$  by Joback  $C_{p,c}^\circ$  GCT
  - 3)  $C_{p,d}^\circ = C_{p,d}^\circ(\text{groups})$  by Joback  $C_{p,d}^\circ$  GCT

The variation of  $C_p^\circ$  with temperature is typically modeled by a cubic in temperature:

$$C_p^\circ = C_{p,a} + C_{p,b} T + C_{p,c} T^2 + C_{p,d} T^3. \quad (11.16)$$

The constants of Equation 11.16 are modeled by group contribution estimation techniques. The seven fundamental properties of the estimation procedure are  $T_{br}$ ,  $T_b$ ,  $P_c$ ,  $C_{p,a}^\circ$ ,  $C_{p,b}^\circ$ ,  $C_{p,c}^\circ$ , and  $C_{p,d}^\circ$ .

#### 11.7.4 Resulting Properties and Groups

The estimation procedures used in the automatic design result in 7 fundamental physical properties:

Table 11.4: Consistent Groups for Automatic Refrigerant Design

$-\text{CH}_3$	$>\text{CH}_2$	$>\text{CH}-$	$>\text{C}<$
$=\text{CH}_2$	$=\text{CH}-$	$=\text{C}<$	$=\text{C}=$
$\equiv\text{CH}$	$\equiv\text{C}-$	$\bar{r}\text{CH}_2\bar{r}$	$\bar{r}>\text{CH}\bar{r}$
$>\text{CH}\bar{r}$	$\bar{r}>\text{C}\bar{r}$	$\bar{r}>\text{C}\bar{r}$	$>\text{C}\bar{r}$
$\bar{r}\text{CH}\bar{r}$	$\bar{r}\text{C}\bar{r}$	$\bar{r}\text{C}\bar{r}$	$=\text{C}\bar{r}$
$-\text{F}$	$-\text{Cl}$	$-\text{Br}$	$-\text{I}$
$-\text{OH}$	$-\text{O}-$	$\bar{r}\text{O}\bar{r}$	$>\text{CO}$
$\bar{r}>\text{CO}$	$-\text{CHO}$	$-\text{COOH}$	$-\text{COO}-$
$=\text{O}$	$-\text{NH}_2$	$>\text{NH}$	$\bar{r}>\text{NH}$
$>\text{N}-$	$=\text{N}-$	$\bar{r}\text{N}\bar{r}$	$-\text{CN}$
$-\text{NO}_2$	$-\text{SH}$	$-\text{S}-$	$\bar{r}\text{S}\bar{r}$

$$T_{b_r} \quad T_b \quad P_c \quad C_{p,a}^o \quad C_{p,b}^o \quad C_{p,c}^o \quad C_{p,d}^o$$

Table 11.4 displays the consistent groups for the group contribution techniques for these properties.

## 11.8 Interactive Design

The estimation procedures developed in Section 11.7.1 for interactive design enable us to design in a two dimensional physical property space. The axes of the space are  $F_1$  and  $F_3$ . Each physical property constraint, being a function of  $F_1$  and  $F_3$  through use

of these estimation procedures, is plotted in this space. This two dimensional physical property space with our physical property constraints is shown in Figure 11.3.

The graphical representation used in the interactive design procedure provides insight into our problem. The first insight is that the constraints we chose yield a feasible solution space. Although I justified each of the physical property constraints there was nothing to indicate that these would result in a feasible space. Additionally, we see that for a major portion of the design space the *Pvp High* constraint is redundant. It is superseded by the *Hv Large* constraint.

Figure 11.4 shows the group vectors for Freon-12. The group vectors indicate that Freon-12 has a high vapor pressure near the constraint. Deviations from the *Cpl Small* and *Hv Large* constraints which were set from Freon-12's values are due to inaccuracies in the estimation procedures.

### 11.8.1 Interactive Results

Table 11.5 shows 19 molecules designed interactively. The comments following each of the compounds were made at the time of the design. These comments refer to the group vectors' satisfaction of the physical property constraints. Figure 11.5 shows the group vectors for one of these designed molecules.

Table 11.6 shows estimated values for the physical properties used in the constraints for the designed refrigerants. Estimates were obtained using the more accurate estimation procedures developed for automatic design. Estimates for fluorinated compounds have greater errors. Table 11.7 shows literature values for some of the designed refrigerants.

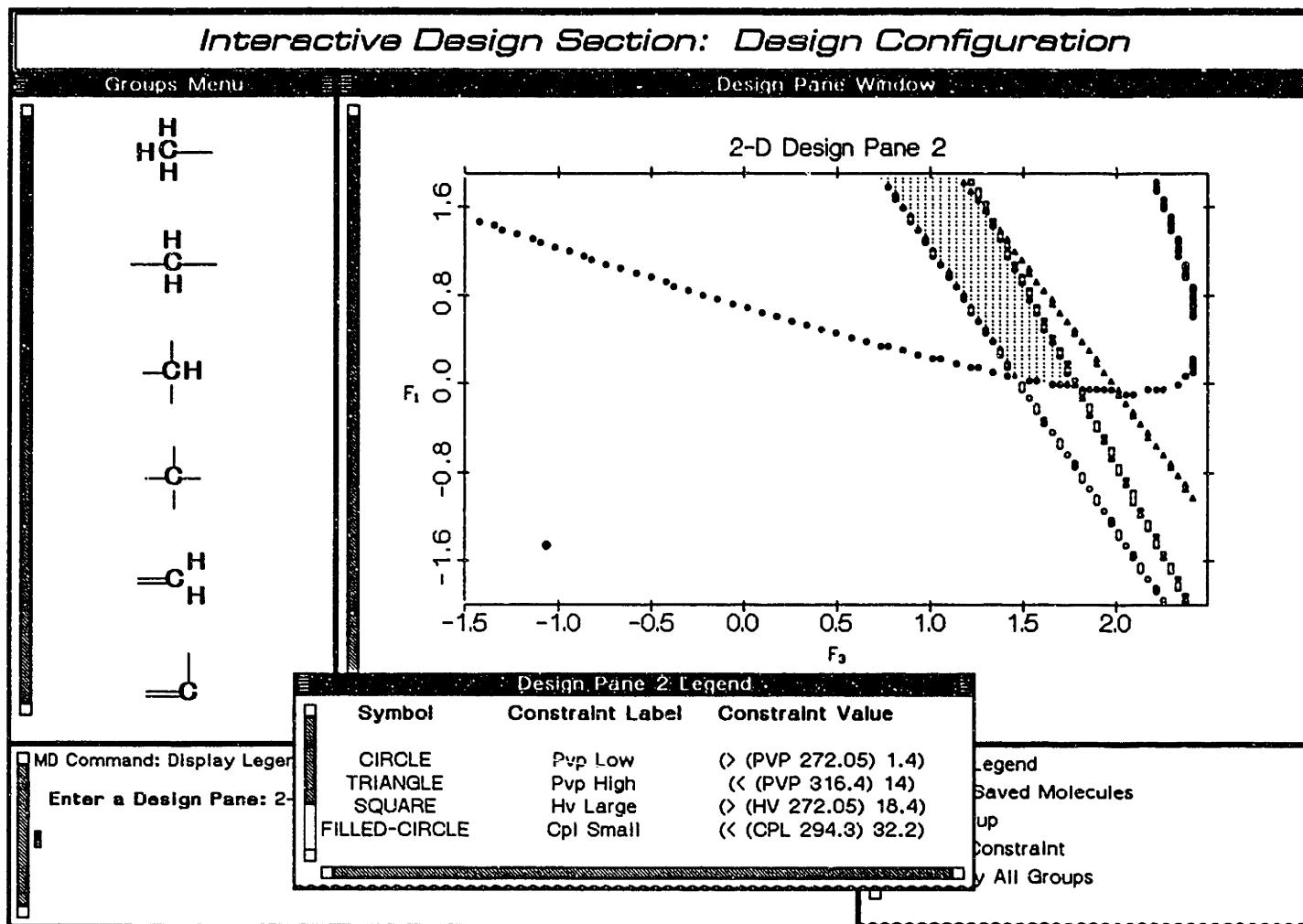


Figure 11.3: Refrigerant Design: Interactive Design Space

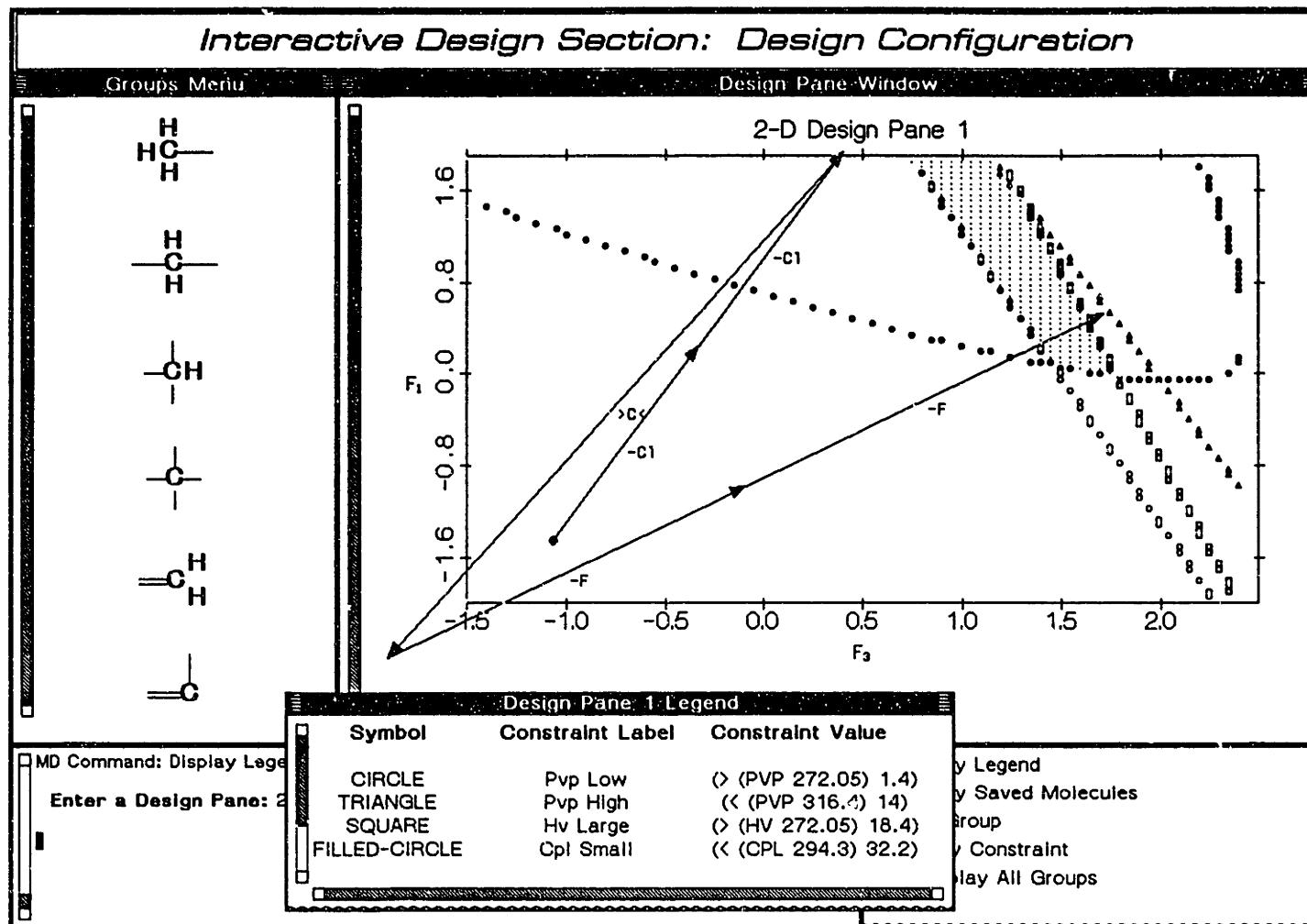


Figure 11.4: Refrigerant Design: Freon 12

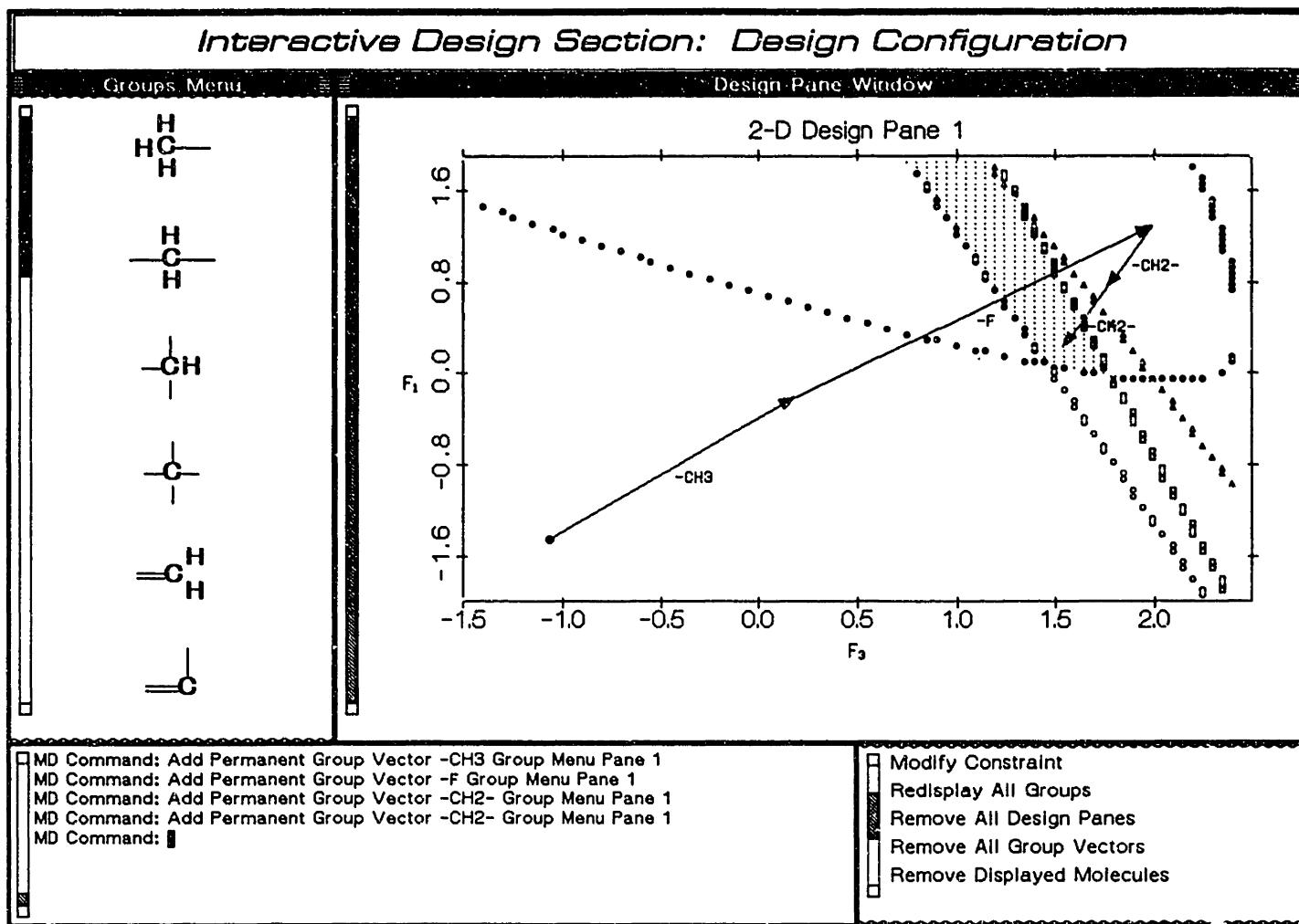


Figure 11.5: Refrigerant Design: Example Molecule

Table 11.5: Designed Refrigerants

	Compound	Design Remarks
1)	CH <sub>3</sub> –CH <sub>3</sub>	Near Hv Large, Near Pvp High
2)	CH <sub>3</sub> –Cl	Near Pvp Low
3)	CH <sub>3</sub> –NH <sub>2</sub>	Center
4)	CH <sub>3</sub> –CH <sub>2</sub> –CH <sub>2</sub> –F	Near Cpl Small
5)	CH <sub>3</sub> –O–O–O–O–F	Near Hv Large
6)	CH <sub>2</sub> =CH–CH <sub>3</sub>	Near Hv Large, Near Pvp High
7)	CH <sub>2</sub> =CH–Cl	Near Pvp Low
8)	CH <sub>2</sub> =CH–O–CH <sub>3</sub>	Center
9)	F–CH <sub>2</sub> –CH <sub>2</sub> –CH <sub>2</sub> –F	Near Pvp High, Outside Hv Large
10)	F–CH=CH–CH <sub>2</sub> –F	Near Hv Large
11)	CCl <sub>2</sub> F <sub>2</sub>	Near Pvp High, Outside Hv Large
12)	CBrClF <sub>2</sub>	Near Pvp Low
13)	CHBrF <sub>2</sub>	Center
14)	CH(COOH)F <sub>2</sub>	Near Pvp Low
15)	CH(HCO)F <sub>2</sub>	Center
16)	NH <sub>2</sub> –NH <sub>2</sub>	Near Pvp Low
17)	CH≡C–Cl	Outside Pvp Low
18)	CH≡C–CH <sub>3</sub>	Center
19)	CH≡C–NH <sub>2</sub>	Near Pvp Low

Table 11.6: Estimated Property Values for Designed Refrigerants

Compound	$P_{vp}$	$P_{vp}$	$\Delta H_v$	$C_{pL}$
	272.05K	316.4K	272.05K	294.3K
1) $\text{CH}_3\text{--CH}_3$	2.710	9.486	18.67	22.49
2) $\text{CH}_3\text{--Cl}$	1.590	6.193	21.58	19.99
3) $\text{CH}_3\text{--NH}_2$	0.375	2.152	29.78	23.50
4) $\text{CH}_3\text{--CH}_2\text{--CH}_2\text{--F}$	1.195	5.013	21.21	31.10
5) $\text{CH}_3\text{--O--O--O--O--F}$	0.478	2.708	25.21	32.88
6) $\text{CH}_2=\text{CH--CH}_3$	1.310	5.195	21.24	25.33
7) $\text{CH}_2=\text{CH--Cl}$	0.750	3.340	24.14	23.01
8) $\text{CH}_2=\text{CH--O--CH}_3$	0.543	2.648	24.83	29.83
9) $\text{F--CH}_2\text{--CH}_2\text{--CH}_2\text{--F}$	1.236	5.330	20.34	34.03
10) $\text{F--CH=CH--CH}_2\text{--F}$	1.045	4.573	20.54	29.89
11) $\text{CCl}_2\text{F}_2$	0.440	2.213	24.70	27.87
12) $\text{CBrClF}_2$	0.122	0.837	28.10	29.02
13) $\text{CHBrF}_2$	0.538	2.855	22.96	26.04
14) $\text{CH}(\text{COOH})\text{F}_2$	0.002	0.041	42.56	37.58
15) $\text{CH}(\text{HCO})\text{F}_2$	0.417	2.421	25.91	30.42
16) $\text{NH}_2\text{--NH}_2$	0.027	0.311	41.41	26.03
17) $\text{CH}\equiv\text{C--Cl}$	0.968	4.084	24.33	21.72
18) $\text{CH}\equiv\text{C--CH}_3$	1.670	6.297	21.44	24.06
19) $\text{CH}\equiv\text{C--NH}_2$	0.214	1.353	32.56	25.42

Table 11.7: Literature Values for some Designed Refrigerants

Compound	$P_{vp}$	$P_{vp}$	$\Delta H_v$	$C_{pL}$
	272.05K	316.4K	272.05K	294.3K
1) $\text{CH}_3\text{--CH}_3$	14.758	27.559	10.81	29.85
2) $\text{CH}_3\text{--Cl}$	1.257	3.128	20.20	19.26
6) $\text{CH}_2=\text{CH--CH}_3$	3.197	6.920	17.01	23.74
11) $\text{CCl}_2\text{F}_2$	1.586	3.705	19.43	28.00

## 11.8.2 Replacing Chlorine

Removing chlorine from current refrigerants would seem to eliminate the hazard of ozone depletion. The interactive design procedure is well suited for searching for group replacements. The group vector for the chlorine group is shown in Figure 11.6. Our target is to find one or more groups contributing similarly to chlorine.

Beginning our search for single group replacements we first restrict the possible groups to those with one single free bond. These groups are then sorted with respect to distance. The three closest groups



are displayed in the design space shown in Figure 11.7.

Again the graphical representation used by the interactive design enables us to evaluate the effect of any substitution. Replacing  $-\text{Cl}$  by  $-\text{Br}$  would result in a compound having reduced vapor pressure, an increased enthalpy of vaporization, and little change in liquid heat capacity. These alterations of physical properties are derived from the relative positions of the group vectors and the locations of the constraints.

## 11.9 Automatic Design

The estimation procedures developed in Section 11.7.3 for the automatic design give us 44 groups with which to search for new refrigerants. Design results indicated that the contributions for the  $=\text{O}$  group made it very desirable for inclusion in candidate molecules. This is unfortunate for two reasons. First, the  $=\text{O}$  group is suggested to be used only when the groups:

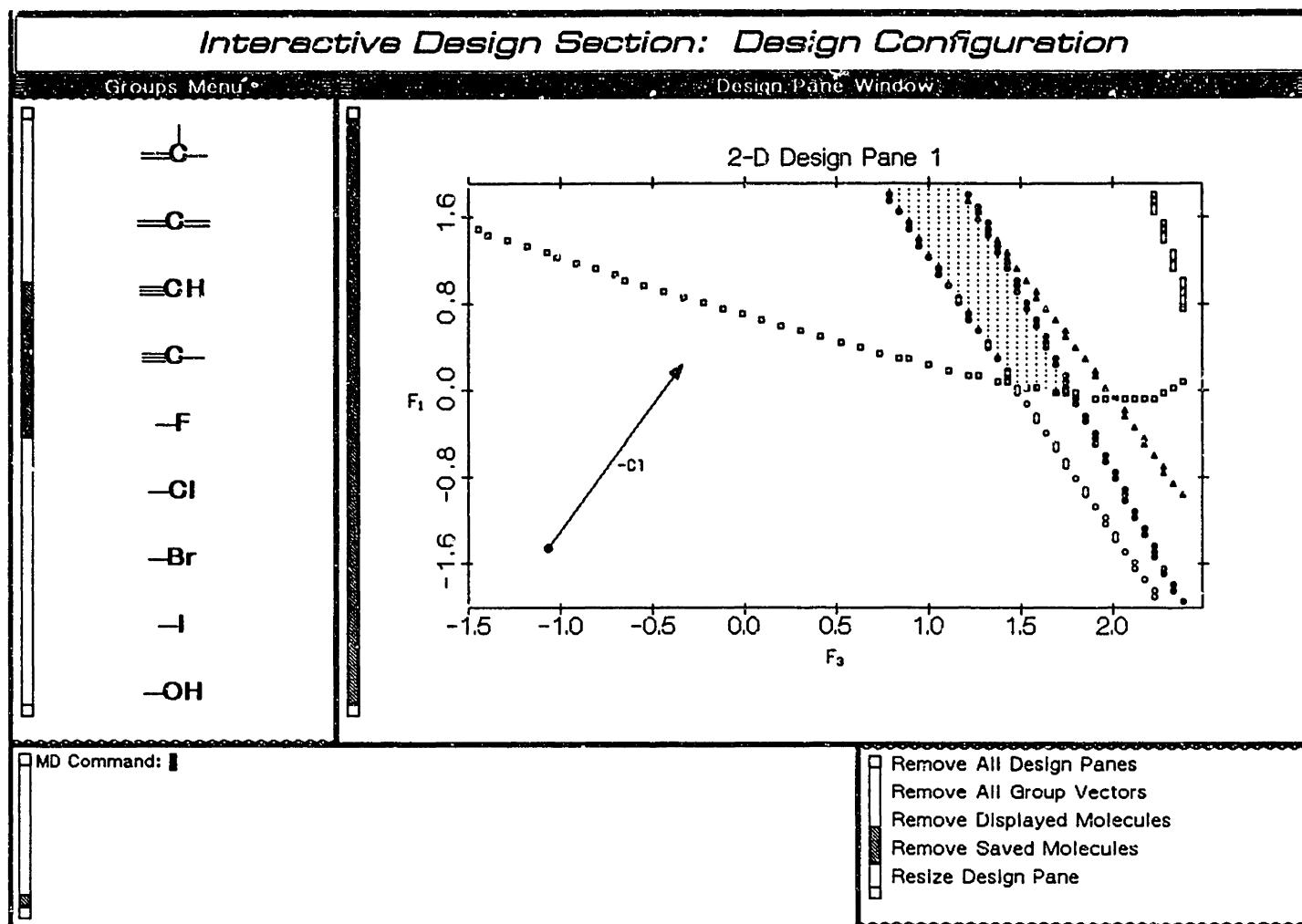


Figure 11.6: Chlorine Group Vector

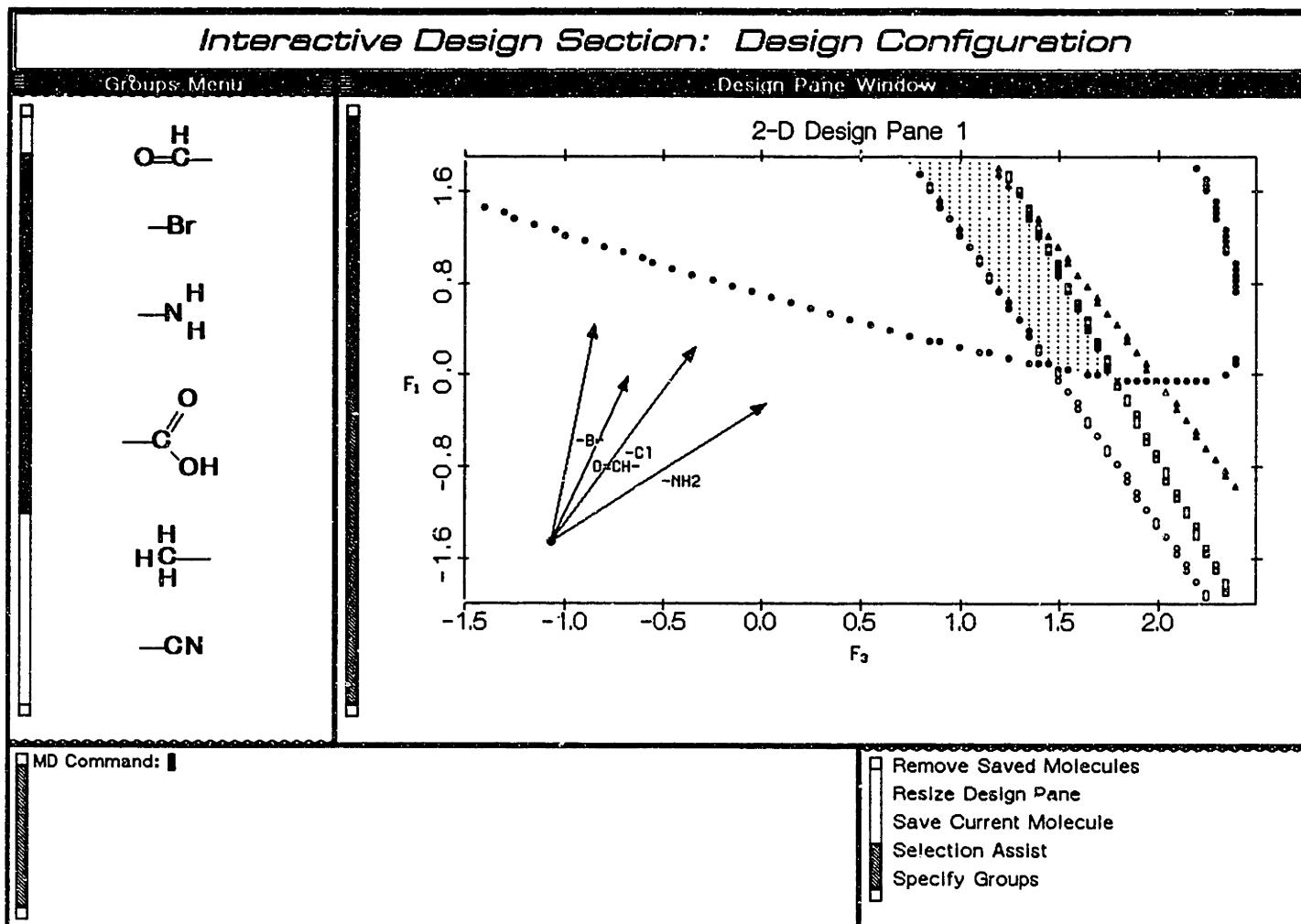


Figure 11.7: Candidate Groups for Chlorine Replacement



are not appropriate. Thus,  $=\text{O}$ 's contribution was derived more from compounds like ketene than ketones and acids. Estimating an aldehyde from a combination of  $-\text{CH}=\text{}$  and  $=\text{O}$  is likely to be inaccurate.

This problem of “group overlap” is common in the automatic design. Because the exact combination of groups is not known until late in the design, restrictions on combinations would be ineffective. The only solution to this problem I see at this time is to use a “prime” set of groups. This ensures that no groups can be formed by a combination of others.

The second reason why the desirability of  $=\text{O}$  groups is unfortunate is because it is unlikely that any compound which has several  $=\text{O}$  groups will be stable. For these two reasons, I restricted the occurrence of the  $=\text{O}$  group to a maximum of one in any molecule.

### 11.9.1 Automatic Results

44 groups were used in the automatic design. Molecules containing between 2 and 7 groups were searched. I limited the number of group occurrences to 7 accounting for the fact that most refrigerants are of small molecular weight. Structural constraints requiring knowledge of bond class or bond valence were not used in the design. These structural constraints were used when the resulting molecules were pruned by hand after the design was completed. 47 molecules were designed which satisfy our four physical property constraints. Table 11.8 shows these 47 molecules with values for  $P_{vp}$ ,

$\Delta H_v$ , and  $C_{p_L}$ .

Molecules 46 and 47 are of particular interest. These are two ringed compounds which possess physical properties which satisfy our constraints. Although the chemical stability of these compounds still needs to be verified I consider it a success that the automatic design was able to design such nonobvious compounds.

Table 11.8: Refrigerant Design – Automatic Results

Molecule	$P_{vp}(272.05)$	$P_{vp}(316.45)$	$H_v(272.05)$	$C_{pL}(294.35)$
1) 1(-CH <sub>3</sub> ) 1(-Cl)	1.59	6.20	21.58	20.00
2) 2(-CH <sub>3</sub> )	2.71	9.50	18.67	22.50
3) 1(-F) 1(-SH)	1.51	6.26	21.03	18.61
4) 1(=O) 1(-Cl) 1(=CH-)	2.43	10.07	26.42	22.28
5) 2(-F) 1(>NH)	2.69	10.96	19.06	22.26
6) 1(-Cl) 1(-F) 1(-CH <sub>2</sub> -)	1.65	6.61	20.71	22.89
7) 1(-Cl) 1(-F) 1(-O-)	1.71	7.06	20.87	21.46
8) 2(=CH <sub>2</sub> ) 1(=C=)	1.53	5.93	20.85	24.92
9) 1(≡CH) 1(-CH <sub>3</sub> ) 1(≡C-)	1.67	6.30	21.44	24.06
10) 1(=O) 1(-F) 1(=CH-) 1(>NH)	1.53	7.75	28.46	27.53
11) 1(=O) 1(-CH <sub>3</sub> ) 1(-CH <sub>2</sub> -) 1(=CH-)	1.71	7.45	27.05	30.46
12) 1(=O) 1(-CH <sub>3</sub> ) 1(-O-) 1(=CH-)	1.77	7.99	27.20	29.07
13) 1(=O) 1(-CH <sub>3</sub> ) 1(=CH-) 1(=C=)	1.49	6.62	27.77	28.76
14) 1(=O) 1(=CH <sub>2</sub> ) 2(=CH-)	1.96	8.25	26.17	27.59
15) 1(=O) 1(=CH <sub>2</sub> ) 2(=C=)	1.78	7.67	27.32	28.34
16) 1(=O) 1(≡CH) 1(=CH-) 1(≡C-)	2.55	10.20	26.26	26.31
17) 2(-F) 2(≡C-)	2.09	7.87	19.61	21.97
18) 1(≡CH) 1(-F) 1(-CH <sub>2</sub> -) 1(≡C-)	1.74	6.72	20.56	26.95
19) 1(≡CH) 1(-F) 1(-O-) 1(≡C-)	1.80	7.18	20.72	25.51
20) 1(=O) 1(-Cl) 1(-F) 1(=C<)	2.54	10.55	25.59	24.86

Table 11.8 Continued: Refrigerant Design – Automatic Results

Molecule	$P_{vp}(272.05)$	$P_{vp}(316.45)$	$H_v(272.05)$	$C_{pL}(294.35)$
21) 1(=O) 2(-CH <sub>3</sub> ) 1(=C<)	1.71	7.29	27.14	30.06
22) 1(-Cl) 2(-F) 1(>N-)	2.65	10.41	19.04	24.83
23) 1(-Cl) 2(-F) 1(>CH-)	1.74	6.96	19.45	25.28
24) 2(-CH <sub>3</sub> ) 1(-F) 1(>N-)	1.81	7.29	20.42	30.04
25) 1(=O) 1(-F) 1(-O-) 1(-CH <sub>2</sub> -) 1(=CH-)	1.86	8.56	26.29	32.08
26) 1(=O) 1(-F) 2(-O-) 1(=CH-)	1.93	9.23	26.43	30.73
27) 1(=O) 1(-F) 1(-CH <sub>2</sub> -) 1(=CH-) 1(=C=)	1.56	7.07	26.88	31.71
28) 1(=O) 1(-F) 1(-O-) 1(=CH-) 1(=C=)	1.61	7.59	27.03	30.34
29) 1(=O) 1(-F) 3(=CH-)	1.50	6.78	26.47	29.26
30) 1(=O) 2(-F) 1(>NH) 1(=C<)	1.59	8.12	27.64	30.11
31) 1(=O) 1(-Cl) 1(-F) 1(=CH-) 1(>N-)	1.53	7.42	28.43	30.03
32) 1(=O) 1(-CH <sub>3</sub> ) 1(-F) 1(=CH-) 1(>N-)	2.83	12.04	25.08	32.54
33) 1(=O) 1(-CH <sub>3</sub> ) 1(-F) 1(-O-) 1(=C<)	1.85	8.37	26.38	31.65
34) 1(=O) 1(-CH <sub>3</sub> ) 1(-F) 1(=C=) 1(=C<)	1.55	6.93	26.96	31.30
35) 1(=O) 1(≡CH) 1(-F) 1(≡C-) 1(=C<)	2.66	10.69	25.43	28.88
36) 3(-F) 1(>NH) 1(>N-)	1.70	8.12	20.88	29.98
37) 1(-CH <sub>3</sub> ) 2(-F) 1(-O-) 1(>N-)	1.96	8.35	19.70	31.60
38) 1(=CH <sub>2</sub> ) 2(-F) 1(=CH-) 1(>N-)	2.15	8.61	18.73	30.15
39) 1(=CH <sub>2</sub> ) 2(-F) 1(-CH <sub>2</sub> -) 1(=C<)	1.41	5.78	19.59	30.78
40) 1(=CH <sub>2</sub> ) 2(-F) 1(-O-) 1(=C<)	1.45	6.13	19.75	29.34

Table 11.8 Continued: Refrigerant Design – Automatic Results

Molecule		$P_{vp}(272.05)$	$P_{vp}(316.45)$	$H_v(272.05)$	$C_{pL}(294.35)$
41)	$1(=CH_2) 2(-F) 1(=CH-) 1(>CH-)$	1.42	5.82	19.12	30.62
42)	$1(\equiv CH) 2(-F) 1(\equiv C-) 1(>N-)$	2.77	10.54	18.90	28.88
43)	$1(\equiv CH) 2(-F) 1(\equiv C-) 1(>CH-)$	1.83	7.07	19.31	29.34
44)	$1(=O) 2(-F) 2(=CH-) 1(=C<)$	1.56	7.10	25.66	31.83
45)	$3(-F) 2(=CH-) 1(>N-)$	1.66	7.13	18.92	31.79
46)	$2(\overset{r}{>}CH-) 1(=C\overset{r}{<}) 1(=O) 2(-F)$	1.53	7.11	26.08	31.93
47)	$3(\overset{r}{>}CH-) 3(-F)$	1.40	5.80	18.67	31.11

# Chapter 12

## Polymer Design

A common problem in polymer science is finding a polymer which meets a number of physical property constraints[26]. In this case study I demonstrate the applicability of my design methodology to the design of polymers with desired properties. I design polymers for use as integrated circuit(IC) encapsulants.

I begin the case study by describing the encapsulant problem. Polymers currently used as encapsulants are described along with their important physical property values. These physical property values are used in the formulation of the problem. Estimation techniques from Van Krevelen[127] and Salame[110] transform the physical property target into a set of constraints on fundamental properties. The automatic design procedure suggests candidate encapsulants. These suggested compounds are discussed.

## 12.1 IC Packaging

Electronic packages are sealed to prevent gross contamination, handling damage, and the entry of corrosive gases[87]. Polymeric coatings are widely used in the electronics industry because of their good physical properties and low cost[45]. Polymers used for semiconductor encapsulation must protect against moisture, chemical agents, wide temperature variations, and mechanical shock. The polymeric material must be able to do this with minimum effect on device parameters over an extended period of time, and be relatively inexpensive and easy to process.

IC packaging or encapsulation is the process of surrounding a microelectronic device in a thick coat of protective polymer or ceramic. To package microelectronic circuitry so that it is useful and functions properly under various environmental conditions, it is essential to select the correct packaging material[37]. Some of the important physical properties of packaging material are[27]:

1. Permeability to water vapor and oxygen at high temperatures.
2. Thermal conductivity.
3. Outgassing in plastics at elevated temperatures and the resulting impact on water vapor permeability.
4. Thermal expansion coefficient and mismatch between expansion coefficients of package and chip interconnect.
5. Resistivity must be high since the package will interconnect leads.

## 12.2 Current Encapsulants

Polyimides have been used in the fabrication of integrated circuit devices for many years because of their ability to form pinhole and crack-free film, excellent thermal stability,

Table 12.1: Some Physical Properties of Polyimides

Physical Property	Value	Units/Notes
Dielectric Constant	2.8–4.0	at 1 kHz
Dissipation Factor	0.002	at 1 kHz
Dielectric Strength	>100	kV/mm
Volume Resistivity	$10^{14}$ – $10^{16}$	ohm-cm
Thermal Decomposition Temperature	405–650	°C
Glass Transition Temperature	200–400	°C
Thermal Coefficient of Expansion	$10^{-6}$ – $10^{-4}$	cm/cm·°C
Thermal Conductivity at 25°C	0.16	W/m·K
Tensile Strength	90–300	MPa
Tensile Modulus	1.3–6.2	GPa

chemical resistance and dielectric properties[17]. A large variety of polyimides are available. Table 12.1 show some of the important physical properties of polyimides used for microelectronic packaging material[17].

### 12.3 Problem Formulation

The first step in my methodology is to identify the target. The following constraints specify what is desired in a good encapsulant:

- $T_g > 400^\circ\text{C}$

The glass transition temperature must be greater than  $400^\circ\text{C}$ . The encapsulant must keep its structural integrity during use. The high temperatures which microelectronic circuits can operate at place a restriction on  $T_g$ .

- $R > 10^{16}$  ohm-cm

The volume resistivity of the solid polymer must be greater than  $10^{16}$  ohm-cm.

Since packaging material makes contact with the metal leads of the microelectronic device it is essential that the compound have a high volume resistivity.

- $\lambda > 0.16 \text{ w/m}\cdot\text{K}$

The thermal conductivity of the solid polymer is desired to be greater than the thermal conductivity of the currently used polyimide. A high thermal conductivity is desirable allowing the microelectronic circuitry to cool more effectively.

- $P(O_2) < 1.0 \text{ cc-mil}/100\text{in}^2/\text{day}/\text{atm}$

Diffusion of oxygen and water through the polymer to the microelectronic circuitry could cause corrosion and is thus undesirable. To establish a value for this physical property constraint I examined polymers used as barriers. Table 12.2 shows the permeability to oxygen for a number of polymers. Polymers with a permeability to oxygen of  $1.0 \text{ cc-mil}/100\text{in}^2/\text{day}/\text{atm}$  or less are considered high barrier materials.

With these constraints enumerated the problem formulation step of the design procedure is now complete. The next step is to identify the estimation techniques which are used to determine physical property values. This identification process is done in the next step: target transformation.

## 12.4 Target Transformation

Four physical properties are used in our polymer design target:

1. Glass Transition Temperature,  $T_g$ .
2. Permeability to Oxygen,  $P(O_2)$ .

Table 12.2: Barrier Polymers

Polymer	$P_{O_2}$ at 25°C cc-mil/100in <sup>2</sup> /day/atm
1 Polybutadiene	3200.
2 Polyethylene	3000.
3 Polycarbonate	300.
4 Polystyrene	350.
5 Tyril 880	72.
6 Polyethylene terephthalate	24.
7 Dow Nitrile resin	12.
8 Polymethyl methacrylate	17.
9 Polyvinyl chloride	7.
10 Dow Nitrile resin	1.5
11 High barrier Saran	1.3
12 Polymethacrylonitrile	0.15
13 Polyacrylonitrile	0.025

3. Thermal Conductivity,  $\lambda$ .

4. Volume Resistivity,  $R$ .

I design polymers using only the automatic procedure. Only one transformed target is needed. In the target transformation step of the methodology it is necessary to identify estimation procedures for each of these physical properties. Estimation techniques from van Krevelen[127] and Salame[110] are used to establish these procedures.

#### 12.4.1 Transformation for Automatic Design

The number of fundamental properties is not important when performing an automatic design. Estimation procedures are developed which yield the greatest accuracy. I developed the following estimation procedures for automatic design.<sup>1</sup>

<sup>1</sup>See Appendix A for a description of the estimation techniques used.

**Thermal Conductivity:** The thermal conductivity is used in one constraint:

$$\lambda > 0.16 \text{ w/m}\cdot\text{K}. \quad (12.1)$$

The estimation procedure for the thermal conductivity follows:

- 1)  $\lambda(298\text{K}) = \lambda(C_p^s, V, U)$  – van Krevelen EOT
- 2)  $C_p^s = C_p^s(\text{groups})$  – van Krevelen GCT
- 2)  $V = V(\text{groups})$  by modified van Krevelen GCT
- 2)  $U = U(\text{groups})$  by van Krevelen GCT

The resulting fundamental properties of the estimation procedure are  $C_p^s$ ,  $V$ , and  $U$ .

**Electrical Resistivity:** The electrical resistivity is used in one constraint:

$$R > 10^{16} \text{ ohm}\cdot\text{cm}. \quad (12.2)$$

The estimation procedure for the volume resistivity follows:

- 1)  $R = R(P_{LL}, V)$  – van Krevelen EOT
- 2)  $P_{LL} = P_{LL}(\text{groups})$  – van Krevelen GCT
- 2)  $V = V(\text{groups})$  – modified van Krevelen GCT

The resulting fundamental properties of the estimation procedure are  $P_{LL}$  and  $V$ .

**Glass Transition Temperature:** The glass transition temperature is used in one constraint:

$$T_g > 400^\circ\text{C}.$$

The estimation procedure for the glass transition temperature follows:

- 1)  $T_g = T_g(Y_g, M)$  by van Krevelen EOT
- 2)  $Y_g = Y_g(\text{groups})$  by van Krevelen GCT
- 2)  $M = M(\text{groups})$  by van Krevelen GCT

The resulting fundamental properties of the estimation procedure are:  $Y_g$  and  $M$ .

**Permeability:** The permeability to oxygen is used in one constraint:

$$P(O_2) < 1.0 \text{cc-mil}/100\text{in}^2/\text{day/atm}. \quad (12.3)$$

The estimation procedure for the permeability to oxygen follows:

- 1)  $P = P(A, S, \pi)$  by Salame EOT
- 2)  $A = 8850 \text{ cc-mil}/100\text{in}^2 \cdot \text{day} \cdot \text{atm}$  for oxygen permeant
- 2)  $S = 0.112$  for oxygen permeant
- 2)  $\pi = \pi(\pi_i, N_b)$  by Salame EOT
- 3)  $\pi_i = \pi_i(\text{groups})$  by Salame GCT
- 3)  $N_b = N_b(\text{groups})$  by Salame GCT

The resulting fundamental properties of the estimation procedure are  $\pi_i$  and  $N_b$ .

### 12.4.2 Consistent Groups

The developed estimation procedures require the seven fundamental physical properties:

$$C_p^s \quad V \quad U \quad P_{LL} \quad Y_g \quad M \quad \pi.$$

The group contribution estimation techniques for each of these fundamental properties has its own set of groups. To perform design it is necessary to find a set of groups consistent with each of these group sets. Table 12.3 shows the 21 consistent groups.

Table 12.3: Consistent Groups for Polymer Design

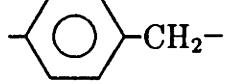
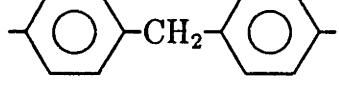
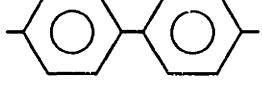
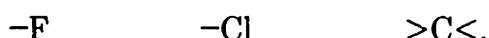
$-\text{CH}_2-$	$-\text{CH}(\text{CH}_3)-$
$-\text{CH}(\text{C}_6\text{H}_5)-$	$-\text{C}(\text{CH}_3)_2-$
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	
 -CH <sub>2</sub> -	$-\text{CH}_2-\text{C}_6\text{H}_5-\text{CH}_2-$
	
-O-	$-\text{COO}-$
$-\text{O}-\text{COO}-$	$-\text{CO}-\text{NH}-$
$-\text{O}-\text{CO}-\text{NH}-$	$-\text{CHF}-$
$-\text{CF}_2-$	$-\text{CHCl}-$
$-\text{CCl}_2-$	$-\text{CH}=\text{CCl}-$
$-\text{CFCl}-$	

Table 12.3 shows good examples of the concept of group inclusion. van Krevelen's group set for  $Y_g$  contains the group:



Salame's group set for  $\pi$  does not contain this group. However, it does contain the groups:



Thus van Krevelen's group is included in Salame's group and is usable.

## 12.5 Automatic Results

The 21 groups of Table 12.3 were used in searching for new polymers satisfying our four physical property constraints. I design polymers having between one and six group occurrences. The design produced a large number of candidate molecules, over 18,000. Table 12.4 shows 50 polymers randomly selected from this candidate set along with estimated values for their important physical properties.

Table 12.4: Polymer Design – Automatic Results

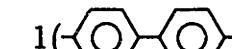
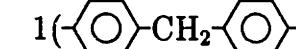
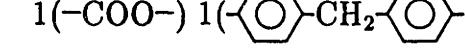
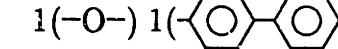
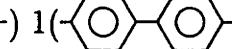
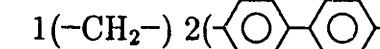
Molecule	$T_g$	$R$	$L$	$Pi$
1) 1(  )	506.0	20.0	0.167	1.29e-02
2) 1(  )	479.5	20.1	0.164	2.40e-03
3) 1(-OCONH-)	423.5	122.6	0.214	3.25e-15
4) 1(-CHCl-) 1(  )	483.4	19.9	0.160	2.52e-02
5) 1(-COO-) 1(  )	417.2	19.8	0.171	1.52e-02
6) 1(-O-) 1(  )	481.6	20.0	0.161	2.12e-01
7) 1(-CONH-) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-)	445.7	19.6	0.172	6.42e-02
8) 1(-CONH-) 1(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 1(-CH(C <sub>6</sub> H <sub>5</sub> )-)	408.8	19.4	0.181	1.74e-02
9) 1(-O-) 1(-CH <sub>2</sub>  1(  )	429.0	20.1	0.161	4.47e-01
10) 1(-CH <sub>2</sub> -) 2(  )	492.2	20.1	0.165	6.49e-01

Table 12.4 Continued: Polymer Design – Automatic Results

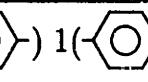
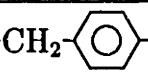
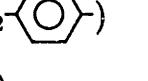
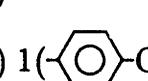
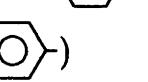
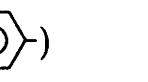
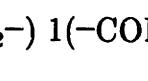
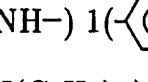
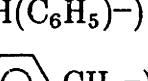
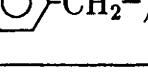
Molecule	$T_g$	$R$	$L$	$Pi$
11) 1(-CH <sub>2</sub> -) 1(  -CH <sub>2</sub> -) 1(  -CH <sub>2</sub> -)	479.5	20.1	0.163	3.71e-01
12) 1(-CH <sub>2</sub> -) 2(  -CH <sub>2</sub> -)	467.9	20.2	0.162	2.12e-01
13) 1(-O-) 2(  -)	493.2	20.0	0.164	8.33e-02
14) 1(-O-) 1(  -CH <sub>2</sub> -) 1(  -CH <sub>2</sub> -)	480.6	20.1	0.162	4.76e-02
15) 1(-O-) 2(  -CH <sub>2</sub> -)	469.0	20.1	0.161	2.72e-02
16) 1(-CHF-) 2(  -)	490.5	20.1	0.163	4.76e-02
17) 1(-CH(CH <sub>3</sub> )-) 1(-CF <sub>2</sub> -) 1(-CONH-) 1(  -)	402.5	19.5	0.160	1.23e-03
18) 1(-CH(CH <sub>3</sub> )-) 1(-CClF-) 1(-CONH-) 1(  -)	414.2	19.3	0.161	6.44e-04
19) 1(-CH(CH <sub>3</sub> )-) 1(-CONH-) 1(-CH(C <sub>6</sub> H <sub>5</sub> )-) 2(  -)	448.9	19.7	0.173	1.20e-02
20) 1(-CHCl-) 1(-CONH-) 2(-CH <sub>2</sub>  -CH <sub>2</sub> -) 1(  -)	406.7	19.7	0.165	9.37e-04

Table 12.4 Continued: Polymer Design – Automatic Results

Molecule	$T_g$	$R$	$L$	$P_i$
21) 1(-CHCl-) 1(-CONH-) 1(-CH <sub>2</sub> -  -) 1(-CH <sub>2</sub> -  -CH <sub>2</sub> -) 1(-  -  -)	402.3	19.6	0.167	1.31e-03
22) 1(-CHCl-) 1(-CONH-) 2(-CH <sub>2</sub> -  -CH <sub>2</sub> -) 1(-  -CH <sub>2</sub> -  -)	400.2	19.7	0.163	6.70e-04
23) 1(-CHCl-) 1(-CONH-) 1(-CH <sub>2</sub> -  -CH <sub>2</sub> -) 2(-  -  -)	446.8	19.6	0.168	4.79e-04
24) 1(-CH <sub>2</sub> -) 1(-CHF-) 1(-CH <sub>2</sub> -  -) 2(-  -  -)	448.3	20.2	0.161	8.12e-01
25) 1(-O-) 1(-CHF-) 1(-CH <sub>2</sub> -  -) 2(-  -  -)	449.3	20.1	0.160	2.37e-01
26) 1(-CH <sub>2</sub> -) 1(-CHF-) 3(-  -  -)	486.9	20.1	0.163	2.96e-01
27) 1(-CH <sub>2</sub> -) 1(-CHF-) 2(-  -  -) 1(-  -CH <sub>2</sub> -  -)	478.9	20.2	0.162	2.12e-01
28) 1(-CH <sub>2</sub> -) 1(-CHF-) 1(-  -  -) 2(-  -CH <sub>2</sub> -  -)	471.3	20.2	0.161	1.51e-01
29) 1(-CH <sub>2</sub> -) 1(-CHF-) 3(-  -CH <sub>2</sub> -  -)	464.1	20.2	0.160	1.08e-01
30) 1(-O-) 1(-CHF-) 3(-  -  -)	487.6	20.1	0.162	8.65e-02

Table 12.4 Continued: Polymer Design – Automatic Results

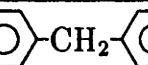
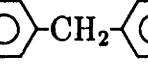
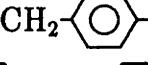
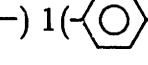
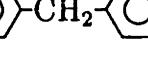
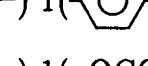
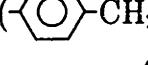
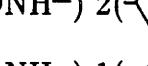
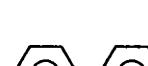
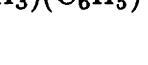
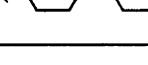
Molecule	$T_g$	$R$	$L$	$P_i$
31) 1(-O-) 1(-CHF-) 2(-  -) 1(-  -CH <sub>2</sub> -  -)	479.6	20.1	0.161	6.18e-02
32) 1(-O-) 1(-CHF-) 1(-  -) 2(-  -CH <sub>2</sub> -  -)	472.0	20.1	0.160	4.42e-02
33) 1(-CH <sub>2</sub> -) 1(-CHF-) 1(-CH <sub>2</sub> -  -CH <sub>2</sub> -) 1(-  -) 1(-OCONH-)	423.6	19.6	0.163	1.75e-03
34) 1(-O-) 1(-CHF-) 2(-  -CH <sub>2</sub> -  -) 1(-OCONH-)	453.7	19.6	0.163	1.33e-04
35) 1(-CH <sub>2</sub> -) 1(-CHF-) 1(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 1(-  -) 1(-OCONH-)	458.5	19.7	0.164	1.20e-02
36) 1(-CH <sub>2</sub> -) 1(-CHF-) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 1(-OCONH-)	442.6	19.8	0.161	1.62e-01
37) 3(-CH <sub>2</sub> -) 1(-CONH-) 2(-  -CH <sub>2</sub> -  -)	429.9	19.8	0.166	7.73e-02
38) 2(-O-) 1(-CH <sub>2</sub> -) 1(-CONH-) 2(-  -CH <sub>2</sub> -  -)	432.0	19.6	0.165	9.92e-03
39) 1(-O-) 2(-CH <sub>2</sub> -) 1(-CONH-) 1(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 1(-  -)	432.1	19.6	0.166	4.23e-01
40) 1(-O-) 1(-CH(CH <sub>3</sub> )-) 4(-  -)	490.2	20.1	0.163	2.33e-01

Table 12.4 Continued: Polymer Design – Automatic Results

Molecule	$T_g$	$R$	$L$	$Pi$
41) 1(-O-) 1(-CH(CH <sub>3</sub> )-) 4(-  -CH <sub>2</sub> -  -)	466.6	20.2	0.160	7.59e-02
42) 1(-CH <sub>2</sub> -) 1(-CHCl- 2(-CH <sub>2</sub> -  -)) 2(-  -  -)	431.9	20.1	0.161	6.14e-01
43) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 3(-  -  -) 1(-OCONH-)	491.8	19.9	0.168	7.78e-03
44) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 2(-CH(C <sub>6</sub> H <sub>5</sub> )-) 1(-  -  -) 1(-OCONH-)	453.5	19.9	0.171	1.60e-01
45) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 3(-  -) 1(-OCONH-)	445.9	19.8	0.168	2.24e-01
46) 1(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 2(-CH(C <sub>6</sub> H <sub>5</sub> )-) 2(-  -) 1(-OCONH-)	421.7	19.8	0.174	1.73e-01
47) 1(-CH <sub>2</sub> -  -CH <sub>2</sub> -) 3(-  -CH <sub>2</sub> -  -) 2(-OCONH-)	453.0	19.5	0.168	6.15e-07
48) 3(-CH(C <sub>6</sub> H <sub>5</sub> )-) 1(-  -  -) 2(-OCONH-)	429.2	19.2	0.182	7.60e-05
49) 3(-CH(C <sub>6</sub> H <sub>5</sub> )-) 1(-  -CH <sub>2</sub> -  -) 2(-OCONH-)	423.2	19.3	0.180	5.74e-05
50) 2(-C(CH <sub>3</sub> )(C <sub>6</sub> H <sub>5</sub> )-) 2(-CH(C <sub>6</sub> H <sub>5</sub> )-) 2(-OCONH-)	434.3	19.4	0.176	1.27e-03

# Chapter 13

## Solvent Design

Liquid-liquid extraction is a technique for separating a solution's components by unequal distribution between two insoluble liquid phases. Liquid extraction is useful when[126]:

- Direct separation methods (e.g., distillation) are more expensive. Separating liquids with poor relative volatilities or requiring high vacuum favors liquid extraction over distillation.
- Direct separation methods are not applicable. Heat sensitive substances and azeotropes can not be separated by distillation.

The success of a liquid-liquid extraction process depends on selecting the most appropriate solvent[77]. In this case study I examine designing a solvent for the extraction of acetic acid from water. I adapted Lo, et.al.'s[77] procedure for solvent selection for use in my interactive design procedure.

### 13.1 The Problem

Recovering acetic acid from aqueous mixtures is an economically important problem in cellulose acetate manufacture, semichemical pulping, RDX manufacture, and other

industries which utilize acetic acid as a raw material or solvent for reaction[77]. Various manufacturing methods for acetic acid involve separation from water. Present-day routes to acetic acid are carbonylation of methanol, liquid-phase oxidation of hydrocarbons such as butane, and oxidation of acetaldehyde. Recently, there has been renewed effort to develop inexpensive routes to acetic acid as a chemical feedstock, based on fermentation of biomass and/or wastes. These also lead to dilute aqueous solutions.

Several important fluid separation processes were initially developed in the context of recovering acetic acid from water. These include liquid-liquid extraction, azeotropic distillation, and extractive distillation. Simple distillation is disadvantageous for two reasons:

1. the relative volatility between water and acetic acid is close to 1.0 and becomes poorest for dilute solutions of acetic acid in water.
2. water is the more volatile component, meaning that for dilute solutions, all the water must be vaporized overhead, leading to a large energy cost per unit of acetic acid recovered.

Brown[15] contrasted liquid extraction with various solvents, azeotropic distillation with benzene or diethyl ketone, and simple distillation. He found extraction most favorable for feeds below 80 weight per cent acetic acid. Eaglesfield[31] concluded that extraction was preferable for feeds below 35 weight per cent acetic acid. Although the thresholds differ both investigations indicate that liquid extraction is the preferred method for separating dilute solutions of acetic acid in water.

## 13.2 Current Solvents

The equilibrium distribution coefficient,  $K_D$ , is one of the most important solvent properties. I express  $K_D$  as the weight fraction of acetic acid in the solvent phase divided by the weight fraction of acetic acid in the aqueous phase at equilibrium. Values of  $K_D$  for several solvents are given in Table 13.1[32].

Table 13.2 summarizes high dilution  $K_D$  data for four homologous series.  $K_D$  undergoes a continual transition between the extreme values shown. In each case the lowest-molecular weight member has the highest value of  $K_D$ . Alcohols have high  $K_D$ 's but tend to esterify with acetic acid and hence are seldom used. Ketones have the next highest  $K_D$ 's but have poor volatility characteristics for solvent recovery by distillation. Acetates and ethers are the more commonly used solvents. For acetates one must guard against loss of solvent due to acid hydrolysis of the ester, but this can usually be avoided with proper precautions.

Eaglesfield et.al.[31] recommends extracting with ethyl acetate for feeds below 16% w/w acetic acid. A mixed solvent composed of ethyl acetate and benzene is recommended for feeds between 16 and 25% w/w acetic acid. The benzene is added to the solvent to suppress coextraction of water, and the amount needed in the solvent mixture increases more or less linearly from zero for the 16% feed to 25% w/w for a 25% feed. Isopropyl acetate has capacity and selectivity properties that offset its lower  $K_D$ . It could be considered as an alternative to ethyl acetate-benzene for feeds above 16% acetic acid. It is the preferred solvent for feeds containing over 25% acetic acid. Brown[15] reached similar conclusions favoring extraction with ethyl acetate and benzene.

Table 13.1: Acetic Acid Distribution Coefficients in Various Solvents

Organic Solvent	°C	$K_D$
Benzene	25	0.04
Toluene	25	0.04
<i>p</i> -Cymene	15	0.02
Diethyl benzene	15	0.016
Diisopropyl benzene	15	0.01
Petroleum ether		0.0001
Kerosene	15	0.003
Dekalin	15	0.015
Tetralin	15	0.01
<i>o</i> -Dichlorobenzene	15	0.03
Ethylene dichloride	15	0.05
Diethyl ether	15	0.50
Diisopropyl ether	23	0.26
Di- <i>n</i> -butyl ether	23	0.10
Diamyl ether	15	0.05
Benzyl ethyl ether	15	0.31
Aniline	15	0.51
Phenol	15	1.40
Cresol	15	0.96
Cyclohexanol	15	0.92
Paraldehyde	18	0.31
Ethyl acetate	18	0.89
Propyl acetate	15	0.42
Butyl acetate	15	0.36
Amyl acetate	15	0.32
Hexyl acetate	15	0.36
Iooctyl acetate	15	0.15
Ethyl aceto-acetate	15	0.85
Phenyl acetate	15	0.29
Benzyl acetate	15	0.25
Cyclohexyl acetate	15	0.47
Methyl cyclohexyl acetate	15	0.48
Ethyl butyrate	15	0.33
Amyl butyrate	15	0.20
Acetophenone	15	0.43
Methyl propyl ketone	20	0.91
Methyl isobutyl ketone	22	0.56
Cyclohexanone	15	1.18
Methyl cyclohexanone	15	0.88

Table 13.2:  $K_D$  for Several Homologous Series

Family	Range of $K_D$
<i>n</i> -Alcohols (C <sub>4</sub> –C <sub>8</sub> )	1.68–0.64
Ketones (C <sub>4</sub> –C <sub>10</sub> )	1.20–0.61
Acetates (C <sub>4</sub> –C <sub>10</sub> )	0.89–0.17
Ethers (C <sub>4</sub> –C <sub>8</sub> )	0.63–0.14

### 13.3 Problem Formulation

The physical properties of the solvent used to facilitate separation in liquid–liquid extraction have major impact on process performance. Two important physical properties are: 1) solvent selectivity for the solute; 2) the solute's partition coefficient between the solvent and parent liquor.

Figure 13.1 shows a simple ternary system representing a liquid–liquid extraction process. S is the solvent used to facilitate separation, A is the parent liquor, and B is the solute which partitions between the solvent and the parent liquor. The partition coefficient for solute B between S and A is given by:

$$m_B = \frac{x_{BS}}{x_{BA}}. \quad (13.1)$$

The selectivity of S for solute B is given by:

$$\beta = \frac{x_{BS}/x_{BA}}{x_{AS}/x_{AA}}. \quad (13.2)$$

At equilibrium the activities of each component is equal in both liquid phases:

$$a_{BS} = a_{BA} \quad a_{AS} = a_{AA} \quad a_{SS} = a_{SA} \quad (13.3)$$

Using the relation:

$$a_{ij} = \gamma_{ij} x_{ij} \quad (13.4)$$

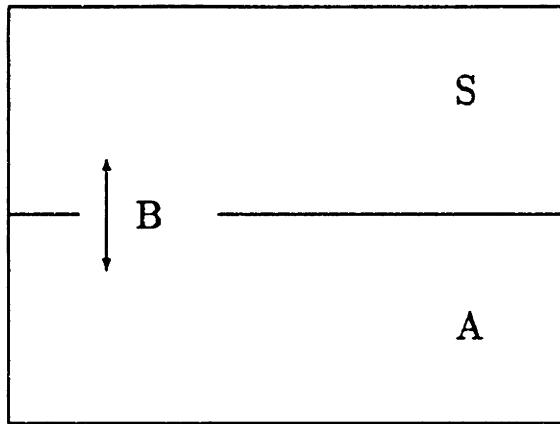


Figure 13.1: Schematic of Solvent Extraction

we express the distribution coefficient and selectivity as:

$$m_B = \frac{\gamma_{BA}}{\gamma_{BS}} \quad \beta = \frac{\gamma_{AS} \gamma_{BA}}{\gamma_{BS} \gamma_{AA}}. \quad (13.5)$$

Assuming that  $\gamma_{AA}$  is 1.0 the selectivity is simplified to:

$$\beta = \frac{\gamma_{AS} \gamma_{BA}}{\gamma_{BS}}. \quad (13.6)$$

## 13.4 Target Transformation

Lo, et. al.[77] used the three component solubility parameter to graphically identify solvents for liquid extraction. The activity coefficient is related to the three component solubility parameter by the equation[77]:

$$\begin{aligned} \ln \gamma_i &= \ln V_i - \ln \bar{V} + \left(1 - \frac{V_i}{\bar{V}}\right) \\ &+ \frac{V_i}{RT} \left\{ (\delta_{D,i} - \bar{\delta}_D)^2 + A_i [(\delta_{p,i} - \bar{\delta}_p)^2 + (\delta_{H,i} - \bar{\delta}_H)^2] \right\} \end{aligned} \quad (13.7)$$

where

$V_i$  = molar volume of component  $i$ .

$\bar{V}$  = mole fraction average molar volume.

$\delta_{D,i}$  = dispersive component of the solubility parameter for component  $i$ .

$\delta_D$  = mole fraction averaged molar volume average of the dispersive component of the solubility parameter.

$\delta_{p,i}$  = dispersive component of the solubility parameter for component  $i$ .

$\bar{\delta}_p$  = mole fraction averaged molar volume average of the polar component of the solubility parameter.

$\delta_{H,i}$  = hydrogen bonding component of the solubility parameter for component  $i$ .

$\bar{\delta}_H$  = mole fraction averaged molar volume average of the hydrogen bonding component of the solubility parameter.

$A_i$  = weighting factor.

To evaluate the partition coefficient and selectivity we require the three activity coefficients:  $\gamma_{AS}$ ,  $\gamma_{BS}$ ,  $\gamma_{BA}$ . The equations for these activity coefficients follow. The superscripts denote either the A or S liquid phases.

$$\begin{aligned}\ln \gamma_{AS} &= \ln V_A - \ln \bar{V}^S + \left(1 - \frac{V_A}{\bar{V}^S}\right) \\ &+ \frac{V_A}{RT} \left\{ (\delta_{D,A} - \bar{\delta}_D^S)^2 + A_A [(\delta_{p,A} - \bar{\delta}_p^S)^2 + (\delta_{H,A} - \bar{\delta}_H^S)^2] \right\} \quad (13.8)\end{aligned}$$

$$\begin{aligned}\ln \gamma_{BS} &= \ln V_B - \ln \bar{V}^S + \left(1 - \frac{V_B}{\bar{V}^S}\right) \\ &+ \frac{V_B}{RT} \left\{ (\delta_{D,B} - \bar{\delta}_D^S)^2 + A_B [(\delta_{p,B} - \bar{\delta}_p^S)^2 + (\delta_{H,B} - \bar{\delta}_H^S)^2] \right\} \quad (13.9)\end{aligned}$$

$$\begin{aligned}\ln \gamma_{BA} &= \ln V_B - \ln \bar{V}^A + \left(1 - \frac{V_B}{\bar{V}^A}\right) \\ &\quad + \frac{V_B}{RT} \left\{ (\delta_{D,B} - \bar{\delta}_D^A)^2 + A_A \left[ (\delta_{p,B} - \bar{\delta}_p^A)^2 + (\delta_{H,B} - \bar{\delta}_H^A)^2 \right] \right\} \quad (13.10)\end{aligned}$$

In systems where mutual solubilities are small, good estimates of distributions may be made by using infinite dilution activity coefficients[77]. Equations 13.8, 13.9, and 13.10 are changed to infinite dilution expressions by substituting the following:

$$\bar{V}^A = V_A$$

$$\bar{V}^S = V_S$$

$$\bar{\delta}^A = \delta_A$$

$$\bar{\delta}^S = \delta_S.$$

For most liquid systems the difference between the dispersive components of the solubility parameter are much smaller than the differences between the polar and hydrogen bonding components[77]. Thus this difference is frequently ignored.

Since components A and B are miscible, their solubility parameters are likely to be nearer to each other than to those of the solvent. Hence[77]:

$$(\delta_B - \bar{\delta}^A)^2 \ll (\delta_B - \bar{\delta}^S)^2 \quad (13.11)$$

$$(\delta_A - \bar{\delta}^A)^2 \ll (\delta_A - \bar{\delta}^S)^2 \quad (13.12)$$

Using the above assumptions, equations 13.8, 13.9, and 13.10 are simplified to:

$$\begin{aligned}\ln \gamma_{AS} &= \ln V_A - \ln V_S + \left(1 - \frac{V_A}{V_S}\right) \\ &\quad + \frac{V_A}{RT} \left\{ A_A \left[ (\delta_{p,A} - \delta_{p,S})^2 + (\delta_{H,A} - \delta_{H,S})^2 \right] \right\} \quad (13.13)\end{aligned}$$

$$\begin{aligned}\ln \gamma_{BS} &= \ln V_B - \ln V_S + \left(1 - \frac{V_B}{V_S}\right) \\ &+ \frac{V_B}{RT} \left\{ A_B \left[ (\delta_{p,B} - \delta_{p,S})^2 + (\delta_{H,B} - \delta_{H,S})^2 \right] \right\} \quad (13.14)\end{aligned}$$

$$\begin{aligned}\ln \gamma_{BA} &= \ln V_B - \ln V_A + \left(1 - \frac{V_B}{V_A}\right) \\ &+ \frac{V_B}{RT} \left\{ A_B \left[ (\delta_{p,B} - \delta_{p,A})^2 + (\delta_{H,B} - \delta_{H,A})^2 \right] \right\} \quad (13.15)\end{aligned}$$

Lo, et.al. separated equations 13.13, 13.9, and 13.10 into two parts. The first contained the solubility parameter terms and the other the molar volume terms. Examining the solubility parameter terms we obtain:

$$m_B \propto \left[ (\delta_{p,B} - \delta_{p,S})^2 + (\delta_{H,B} - \delta_{H,S})^2 \right]^{-1} \quad (13.16)$$

$$\beta_B \propto \frac{(\delta_{p,A} - \delta_{p,S})^2 + (\delta_{H,A} - \delta_{H,S})^2}{(\delta_{p,B} - \delta_{p,S})^2 + (\delta_{H,B} - \delta_{H,S})^2} \quad (13.17)$$

Figure 13.2 shows a plot of the polar and hydrogen bonding components of the solubility parameters with components A, B, and S displayed in the space. In this  $\delta_p/\delta_H$  plane, Equations 13.16 and 13.17 represent distances between points on the plane:

$$m_B \propto r_{B,S}^{-2} \quad (13.18)$$

$$\beta_B \propto \left( \frac{r_{A,S}}{r_{B,S}} \right)^2 \quad (13.19)$$

where  $r_{i,j}$  is the distance between points  $i$  and  $j$  as shown in Figure 13.3.

Equation 13.19 shows that the maximum selectivity is obtained with the maximum distance between components A and S and the minimum distance between components B and S. The best solvent thus lies on a line connecting components B and A closer to component B than to component A. The target region for our design would be

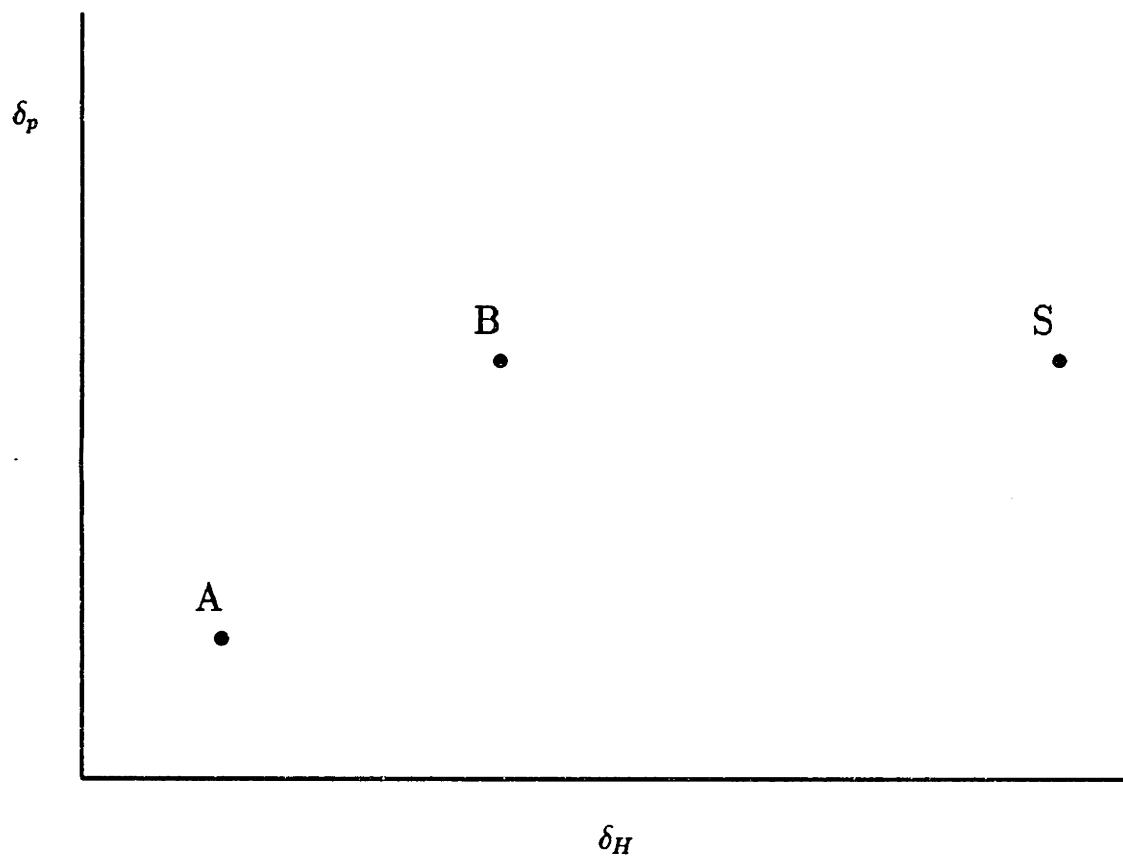


Figure 13.2: Example  $\delta_p$  vs.  $\delta_H$  Parameter Space

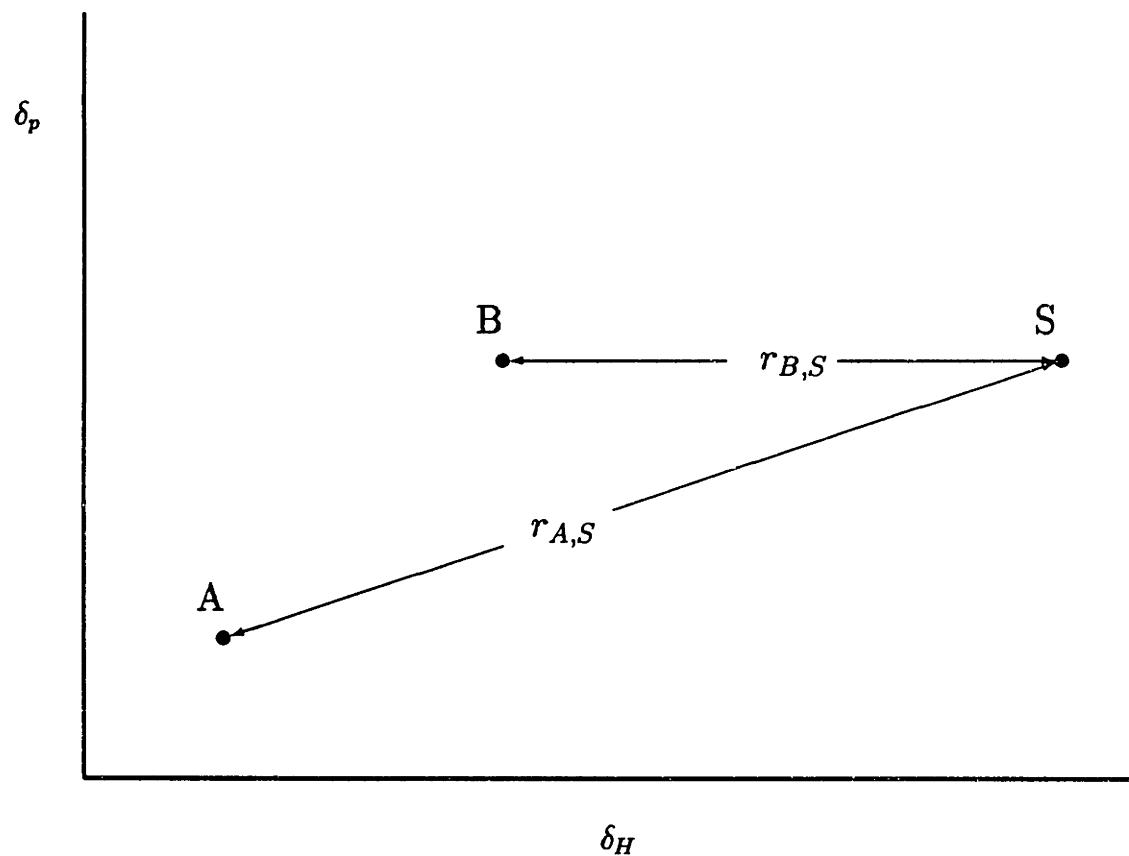


Figure 13.3: Example  $\delta_p$  vs.  $\delta_H$  Parameter Space

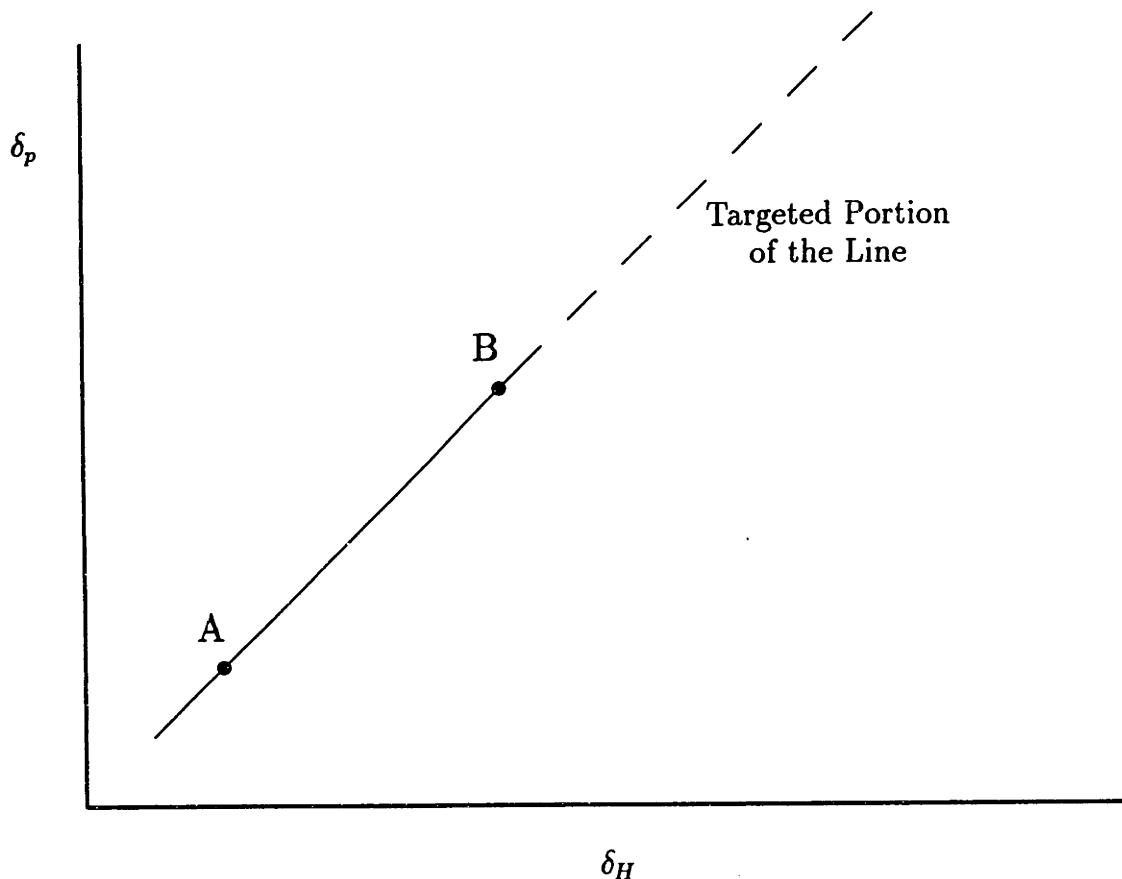


Figure 13.4:  $\delta_p$  vs.  $\delta_H$  Design Space

the portion of a line connecting components A and B on the side of B opposite of A.

Figure 13.4 shows a design space with the target portion of the line dashed.

### 13.4.1 Group Contributions

The approach to solvent selection taken by Lo et.al. is perfectly suited for interactive design. We still need group contribution estimation techniques for  $\delta_p$  and  $\delta_H$ . The published group contribution estimation techniques[6] do not estimate  $\delta_p$  or  $\delta_H$  but estimate  $V\delta_p$  and  $V\delta_H$  where  $V$  is the molar volume of the solvent. In addition, although the estimation technique for  $V\delta_p$  is linear, the estimation technique for  $V\delta_H$

is not.

Using solubility parameter data[6] I developed linear group contribution estimation techniques for both  $\delta_p$  and  $\delta_H$ . These are presented in Appendix A. The estimation techniques were derived from 77 compounds yielding contributions for 15 groups. The average absolute error of the  $\delta_p$  regression was 0.573 with a standard deviation of 0.607. The average absolute error of the  $\delta_H$  regression was 1.08 with a standard deviation of 1.074. Table A.24 shows  $\delta_p$  estimates for 25 compounds not used in the regression. Table A.26 shows  $\delta_H$  estimates for 24 compounds not used in the regression.

The developed estimation techniques show considerable error at times. I believe that more effort would result in improved estimation techniques. However, I consider the developed estimation techniques sufficiently accurate for design purposes and more than adequate to illustrate my solvent design procedure.

### 13.5 Interactive Design

The objective of our interactive design is to design a solvent located on a line passing through the solute and the mother liquor which is close to the solute but far from the mother liquor. The solute in our problem is acetic acid and the mother liquor is water. Table 13.3 shows the polar and hydrogen bonding solubility parameters for acetic acid and water[6].

Figure 13.5 shows our design space. Our target area is the lower left hand portion of the acetic acid–water line. Figure 13.6 shows the group vectors for acetone near our target region. Table 13.4 shows several solvents which were interactively designed

### Interactive Design Section: Design Configuration

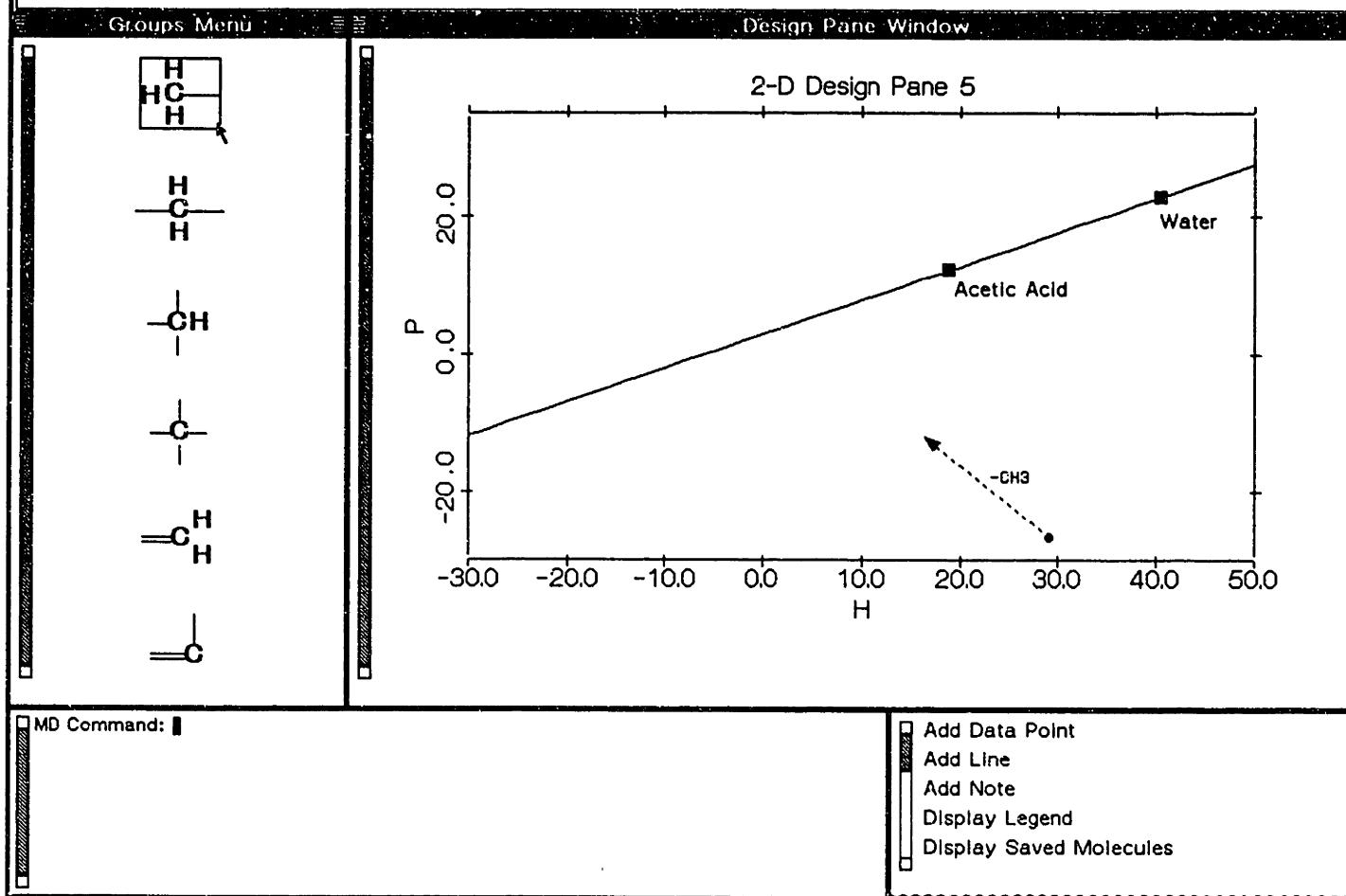


Figure 13.5: Solubility Parameter Design Space

*Interactive Design Section: Design Configuration*

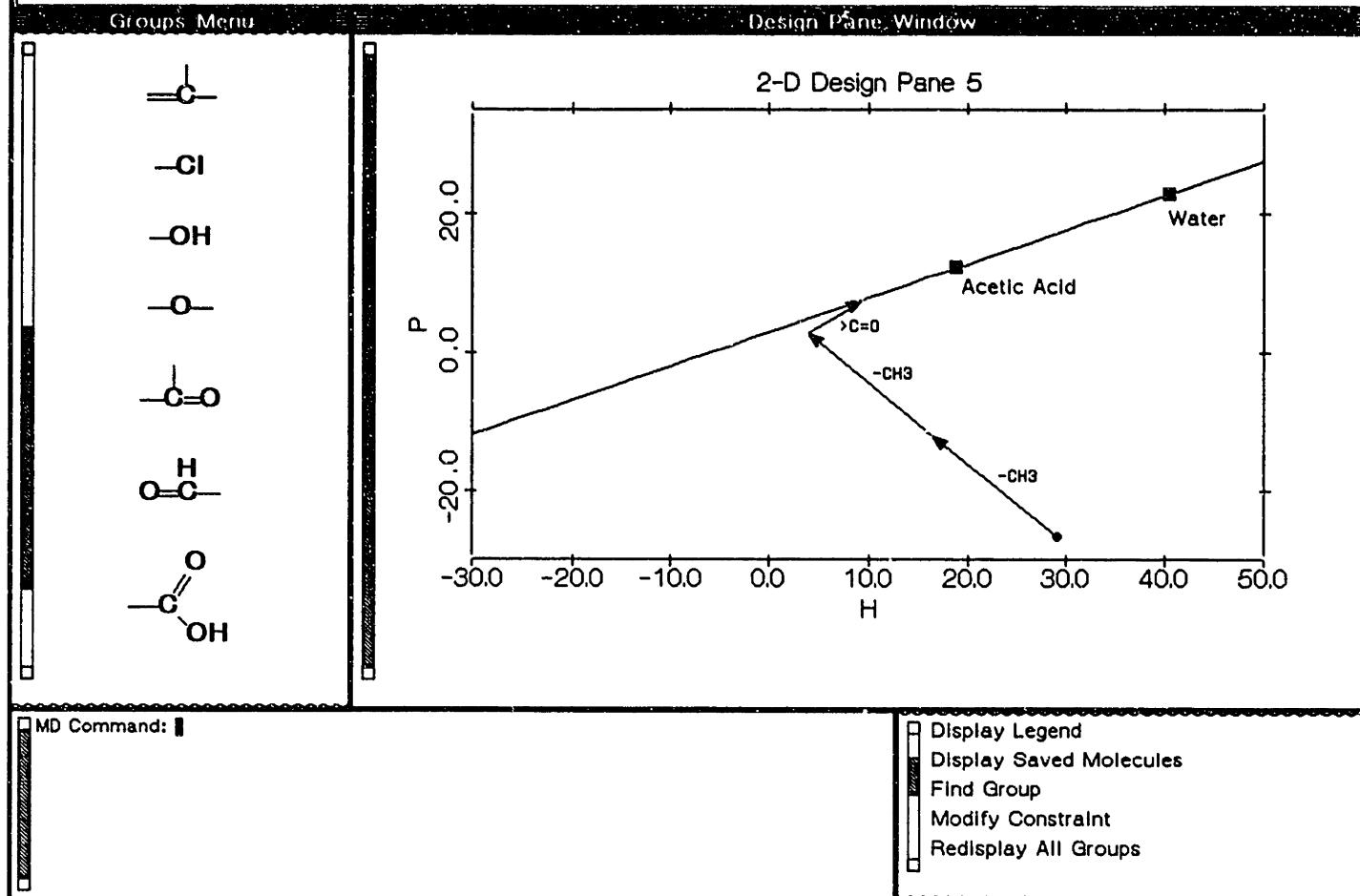


Figure 13.6: Example Solvent

Table 13.3: Acetic Acid and Water Solubility Parameters

	Water	Acetic Acid
$\delta_P$	22.8	12.2
$\delta_H$	40.4	18.9

along with their estimated solubility parameters.

### 13.5.1 Homologous Series

The  $-\text{CH}_2-$  group has a  $\delta_p$  contribution of  $-0.328$  and a  $\delta_H$  contribution of  $-0.512$ .

The slope of  $-\text{CH}_2-$ 's group vector in our interactive design space is  $0.64$ . This is very close to the acetic acid–water target line's slope of  $0.49$ . Adding several  $-\text{CH}_2-$ 's to any of the solvents shown in Table 13.4 thus results in an acceptable new solvent. The direction of the  $-\text{CH}_2-$  group vector is toward lower  $\delta_p$  and  $\delta_H$  values. This reduces the distribution coefficient with increasing  $-\text{CH}_2-$  occurrences. This result is in accordance with the observation reported by Lo et.al.[77] presented in Table 13.2.

### 13.5.2 Solvent Mixtures

Linear mixing rules for solubility parameters enable us to design solvent mixtures. In our  $\delta_p$ – $\delta_H$  design space a binary mixture's solubility parameters lie on a straight line joining the components' solubility parameters. Figure 13.7 shows two solvent pairs,  $S_1$ – $S_2$  and  $S_3$ – $S_4$ , which can be mixed to give good solvent characteristics.

Figure 13.7 shows the solvent pair  $S_3$ – $S_4$  produces a mixed solvent with appropriate solubility parameters. However,  $S_4$  will be selectively extracted from the mixture by

Table 13.4: Liquid-Liquid Extraction Solvents

Solvent	$\delta_p$	$\delta_H$
$\text{CH}_3\text{-Cl}$	7.2	6.2
$\text{CH}_2=\text{CH-CH}_3$	4.7	4.0
$\text{CH}_2=\text{CH-Cl}$	9.2	6.3
$\text{CH}_3\text{-O-CH}_3$	3.8	5.7
$\text{CH}_3\text{-CO-CH}_3$	7.5	9.5
$\text{CH}_2=\text{C(CH}_3)_2$	4.6	4.1
$\text{C(CH}_3)_4$	0.5	1.2
$\text{CCl}_2(\text{CH}_3)_2$	9.5	5.8

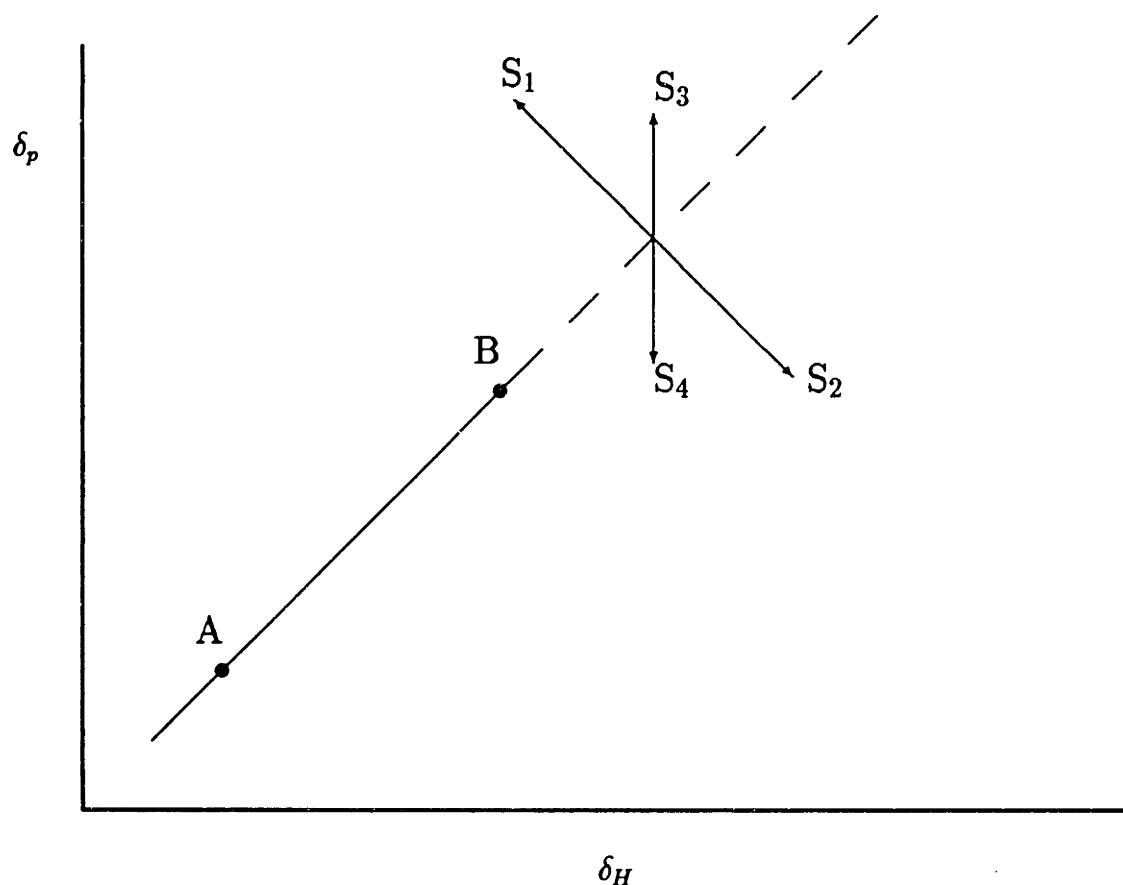


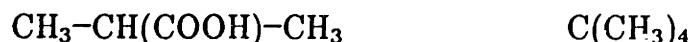
Figure 13.7: Solvent Mixtures

Table 13.5: Solvent Mixtures

Component A	Component B
$\text{CH}_3\text{--CH(OH)--CH}_3$	$\text{CH}_3\text{--}(\text{CH}_2)_8\text{--CH}_3$
$\delta_p = 9.7$	$\delta_p = 0.1$
$\delta_H = -4.9$	$\delta_H = -0.2$
$x = 0.225$	$x = 0.775$
	$\delta_{p,mix} = 2.3$
	$\delta_{H,mix} = -1.3$
$\text{CH}_3\text{--CH(COOH)--CH}_3$	$\text{C}(\text{CH}_3)_4$
$\delta_p = 10.6$	$\delta_p = 0.5$
$\delta_H = -5.2$	$\delta_H = 1.2$
$x = 0.227$	$x = 0.773$
	$\delta_{p,mix} = 2.8$
	$\delta_{H,mix} = -0.3$

the A-B mixture. The solvent pair  $S_1-S_2$  is better since demixing is less likely. Solvent pairs used in mixtures should be equidistant from the solute-parent liquor solubility parameters. Table 13.5 lists two solvent pairs designed for extracting acetic acid from water. The estimated solubility parameters of the pure components, mixture, and composition of the mixture are shown. The negative solubility parameters indicate the error in the estimation techniques.

Figure 13.8 shows the group vectors for the



solvent pair.

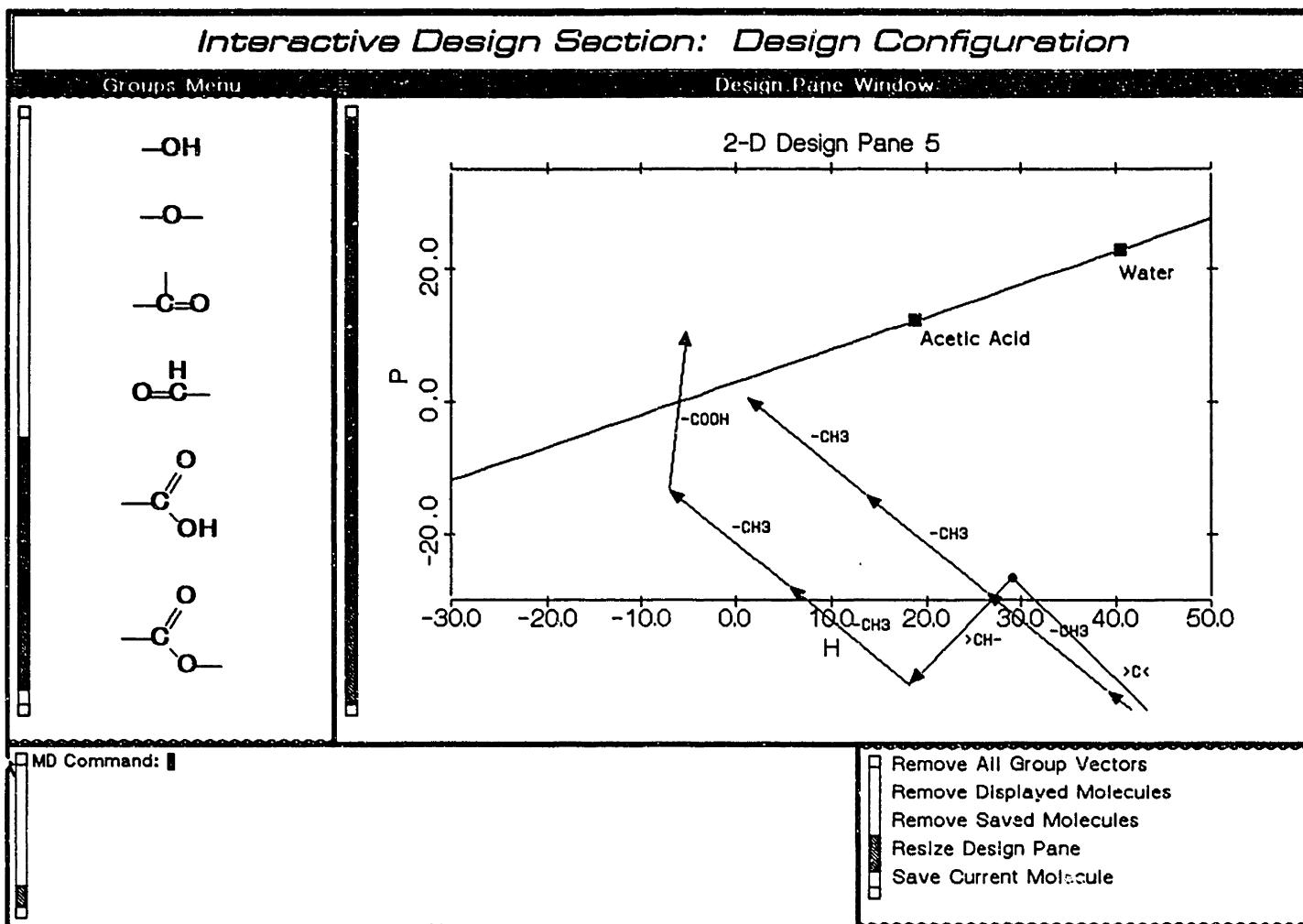


Figure 13.8: Interactive Design of Solvent Mixtures

---

 Table 13.6: UNIFAC Groups
 

---

$-\text{CH}_3$	$-\text{CH}_2-$	$>\text{CH}-$	$>\text{C}<$
$\text{CH}_2=\text{CH}-$	$-\text{CH}=\text{CH}-$	$-\text{CH}=\text{C}<$	$\text{CH}_2=\text{C}<$
$-\text{CH}_2-\text{OH}$	$-\text{CH}(\text{OH})-$	$\text{CH}_3-\text{CO}-$	$-\text{CH}_2-\text{CO}-$
$-\text{CH}_2-\text{CN}$			

---

## 13.6 Automatic Design using UNIFAC

UNIFAC is a group contribution technique for estimating activity coefficients. It is highly nonlinear and incorporates group interactions. I believe that the UNIFAC model can be used in an automatic design.

Table 13.6 shows a set of 16 groups for which UNIFAC group contributions are available. Table 13.7 shows the interaction parameters for classes of groups[80].

An automatic design would begin by collecting the groups into a single meta-group. The  $R$  and  $Q$  UNIFAC parameters would be calculated for this meta-group as any meta-contribution. The interaction parameters would need to be collected for all possible main group interactions. The interval of these values would produce the interaction parameter for the meta-group. For the four meta-groups:

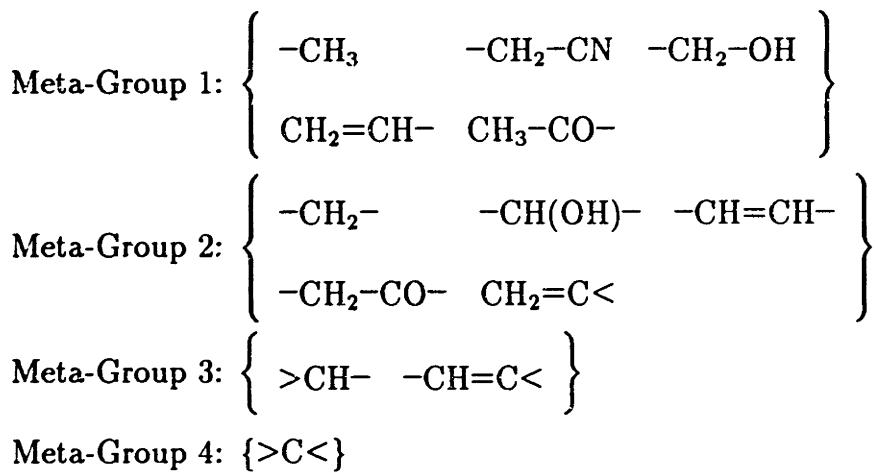
---

 Table 13.7: Some UNIFAC Interaction Parameters
 

---

	$\text{CH}_2$	$\text{C}=\text{C}$	$\text{OH}$	$\text{CH}_2\text{CO}$	$\text{CCN}$
$\text{CH}_2$	0	74.54	644.6	472.6	696.8
$\text{C}=\text{C}$	292.3	0	724.4	343.7	405.9
$\text{OH}$	328.2	470.7	0	67.07	-189.3
$\text{CH}_2\text{CO}$	66.56	306.1	216.0	0	430.6
$\text{CCN}$	29.08	34.78	2011.0	-349.2	0

---



the interaction parameters would be:

	MG-1	MG-2	MG-3	MG-4
MG-1	[-349.2 2011.0]	[-348.2 2011.0]	[0 470.7]	[0 328.2]
MG-2	[-189.3 724.4]	[0 724.4]	[0 470.7]	[0 328.2]
MG-3	[0 724.4]	[0 724.4]	[0 292.3]	[0 292.3]
MG-4	[0 696.8]	[0 644.6]	[0 74.54]	[0 0]

The one additional parameter must be entered before we are able to calculate the activity coefficient. This parameter is the mole fraction. Since the actual mole fraction is not a typical design parameter I believe that it should be entered as an interval. Allowing the concentration to vary over the entire [0 1] range might result in an infinite number of possible solutions. I believe that any design should be done in steps, examining a narrow range of the mole fraction.

The complexity of the UNIFAC model could pose a problem to interval evaluation. The numerous occurrences of mole fractions and group parameters could contribute to excess interval width. Interval tightening techniques such as monotonicity identification will need to be examined.

# Chapter 14

## Drug Design

The QSAR<sup>1</sup> approach to drug design takes a two step approach to the correlation of biological activity with structure. The first step correlates a measure of biological activity, usually the reciprocal concentration having a 50% effect, with a number of physical properties. This correlation yields a model which is used to determine the optimal value of the physical properties giving the maximum potency. These physical properties are determined by group contribution estimation techniques. The second step of the approach is to identify group substitutions which lead to structures possessing these optimal physical property values.

In this chapter I discuss the QSAR approach to drug design. I present an example taken from work done by Cramer et.al.[19,22] demonstrating how the interactive design procedure assists in the second step of the QSAR approach to drug design.

---

<sup>1</sup>Quantitative Structure Activity Relationship

## 14.1 Steps in Drug Design

There are primarily five steps in drug design[84]:

1. Identification of targeted biological properties.
2. Identification of a “lead” compound.
3. Synthesis and testing of a set of “analogs”.
4. Construction of a quantitative structure activity relationship.
5. Generation of compounds which possess physical property values which optimize drug potency.

I briefly describe each of these steps.

### 14.1.1 Target Biological Properties

Step 1 in drug design is to identify the target properties. It is customary in pharmacological experiments to define biological response,  $W$ , as the magnitude of the change in a measurable biological parameter after administration of a drug (e.g. muscle contraction, change in blood pressure, etc.) and to investigate this quantity as a function of drug concentration (dose)[39]. The biological activities of drugs are compared by considering the concentration (dose),  $C$ , which just produces a given response. In this way variables such as  $ED_x$  (dose producing  $x\%$  of maximal activity),  $LD_x$  (lethal dose for  $x\%$  of test objects) and  $I_x$  (dose causing an  $x\%$  inhibition of a function) are obtained[39]. The biological activity is then simply defined as the reciprocal of that dose.

### **14.1.2 Identification of “Lead” Compounds**

Step 2 in drug design is identifying a “lead” structure[84]. This is usually a molecule with demonstrable but weak activity. The search for new lead compounds proceeds in several ways[120]:

1. Isolation, purification, and identification of compounds from natural products, including plant sources, animal sources, and microorganisms. Examples of drugs found in this way include antibiotics, alkaloids, steroids, and cardiac glycosides.
2. Following up leads generated by therapeutic folklore or folk medicine.
3. Testing of metabolites or molecular modifications of metabolites of known drug compounds.
4. Fundamental studies of biochemical systems.
5. The investigation of side effects of experimental or clinically used drugs.
6. Mass screening of chemical compounds for possible biological activity.
7. Organic synthesis aimed at the production of bioactive compounds.

### **14.1.3 Analog Synthesis and Model Development**

Steps 2 and 4 in drug design attempt to develop an understanding of the manner in which a compound’s physical properties affect its biological activity. The physical properties used in most QSAR based drug designs attempt to represent electronic, hydrophobic, and steric effects.

#### **Electronic Effects**

Hammett[49] studied the effect sterically remote substituents had on the equilibrium or rate constants of organic reactions. Substituent effects in such reactions are electronic

in nature[84]. Hammett found that the rates of a substituted compound were linearly related to the rates of the parent compound by:

$$\log k - \log k_0 = \rho\sigma \quad (14.1)$$

in which  $k$  refers to the rate or equilibrium constant for the reaction of the substituted compound,  $k_0$  the rate or equilibrium constant for the parent compound,  $\rho$  is a proportionality constant dependent upon the specific reaction and reaction conditions, and  $\sigma$  is a parameter dependent only upon the substituent.  $\sigma$  is called the Hammett constant.

Substituents with positive  $\sigma$  values are electron withdrawing and those with negative  $\sigma$  values are electron donating. Reactions characterized by positive  $\rho$  values are aided by electron-withdrawing substituents on the benzene ring, and reactions with negative  $\rho$  values are aided by electron-donating substituents[120].

### Hydrophobic Effects

Hansch[50] modeled the hydrophobic properties using the logarithm of the partition coefficient,  $P$ , between a lipid model system, normally n-octanol, and water[120]. The partition coefficient is used to develop an index,  $\pi$ , that expresses the difference between a substituted compound and the parent compound.  $\pi$  is defined by:

$$\pi = \log P_x - \log P_0 \quad (14.2)$$

where  $P_x$  is the partition coefficient for the substituted compound and  $P_0$  is the partition coefficient for the parent compound.

## Steric Effects

Steric effects of substituents on reaction rates are most widely characterized by the Taft[122] parameter,  $E_s$ [120]. Of all the effects for which one must select parameter values, those for steric effects present the most problem[84]. The reason for this is that steric interactions are not easily translated from one compound or reaction type to another, and that it is not always clear which aspect of steric effects might be important.

$E_s$  is defined as the logarithm of the relative rate, under identical conditions of solvent, temperature and acidity, of the acid-catalyzed hydrolysis of the substituted methyl ester compared to that of methyl acetate:

$$E_{sX} = \log k_{XCO_2CH_3} - \log k_{CH_3CO_2CH_3} \quad (14.3)$$

where  $k_{XCO_2CH_3}$  is the substituted ester's rate constant for hydrolysis and  $k_{CH_3CO_2CH_3}$  is the rate constant for hydrolysis of methyl acetate.

## Models

The relationship between biological activity and the physical parameters representing electronic, hydrophobic, and steric effects is often modeled by:

$$\log \frac{1}{C} = k_1\pi + \rho\sigma + k_3E_s + \text{constant.} \quad (14.4)$$

At times it is necessary to include higher order terms in  $\pi$ . Equation 14.4 thus becomes:

$$\log \frac{1}{C} = k_1\pi + k_2\pi^2 + \rho\sigma + k_3E_s + \text{constant.} \quad (14.5)$$

In numerous QSAR studies the steric parameter,  $E_s$ , is not included.

The coefficients of Equation 14.5,  $k_1$ ,  $k_2$ ,  $k_3$ , and  $\rho$ , are determined by regression analysis on a set of known compounds.

### **Selection of Analogs**

To identify values for the coefficients of Equation 14.5 it is necessary to develop a set of data on which to perform regression. The data consists of a set of compounds for which biological activity is measured. The physical parameters used in the regression are then calculated for each compound. A multiple regression is performed on this data yielding the final model.

To ensure a powerful predictive regression model it is necessary to analyze a set of analogs which have a wide distribution of values for each of the physical parameters used in Equation 14.5. It is not easy to hypothesize in an *a priori* QSAR analysis which properties will be important for biological activity. A good strategy to apply in such cases is always to start with pilot studies of small sets of compounds in an attempt to derive a very simple model[39]. The variance of the coefficients for each of the terms in the regression model is then analyzed. Those terms whose coefficients have high variance are analyzed to see if a sufficient sampling of values was incorporated into the data.

#### **14.1.4 Optimization of Drug Potency**

With the model of biological activity developed in step four, the final step in drug design is to determine the values of the physical parameters which give an optimal value of drug potency. Thus, if the coefficient for  $\sigma$  in the developed model was negative compounds

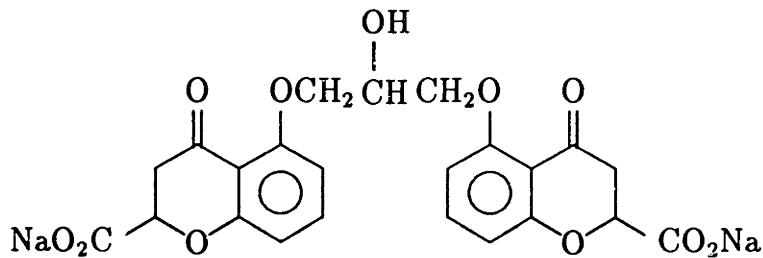


Figure 14.1: Existing Antiallergic Sodium Chromoglycate

with large negative values for  $\sigma$  would have greater potency. Once the desired values for  $\sigma$ ,  $\pi$ , and  $E$ , are identified the process is one of selecting the appropriate substituents which give these values.

This is the step at which the current drug design methodology ends. Procedures for selecting the appropriate substituents have not been incorporated into the methodology. Hansch[52] has suggested the use of cluster analysis to reduce the large number of possible substituents to a more manageable number. However, with the inclusion of more than one substituent, this approach is still overwhelmed by the combinatorics.

The interactive design is a very effective tool for assisting in the selection of appropriate substituents. I demonstrate this in the following sections.

## 14.2 The Problem

Sodium chromoglycate, Figure 14.1, is effective at warding off asthmatic attacks. However, it must be administered by inhalation. Cramer et.al.[19,22] performed a QSAR study to find a more potent pharmaceutical which could be administered orally.

Searching through a database of approximately one thousand compounds they se-

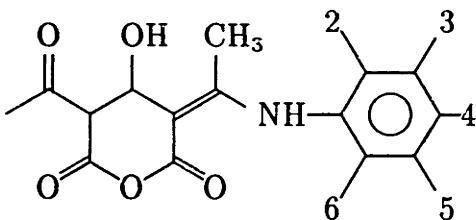


Figure 14.2: “Lead” Compound for Drug Design

lected the pyranenamines, Figure 14.2, as their lead compound. The pyranenamines have biological properties promising enough to merit a synthetic search for structurally related compounds having improved properties. The task was to develop highly active compounds by varying phenyl ring substituents.

### 14.3 Problem Formulation

To relate  $pI_{50}$  to molecular structure Cramer, et.al. followed the two step QSAR approach. Nineteen compounds were synthesized with varying substituents.  $pI_{50}$  was measured for each analog. Values for  $\pi$  and  $\sigma$  were computed for each of the substituent sets. Table 14.1 shows the data gathered[19].

This data was regressed to develop a relationship between  $pI_{50}$  and the biochemical parameters  $\pi$  and  $\sigma$ . The model developed was:

$$pI_{50} = -0.72 - 0.14 \sum \pi - 1.35 \left( \sum \sigma \right)^2 \quad (14.6)$$

Table 14.2 compares model and experimental results for 10 highly active compounds as reported by Cramer. The model shows considerable error but does follow a general trend.

Table 14.1: Data for Initial 19 Analogs

Substituents	$pI_{50}$	$\sum \pi$	$(\sum \sigma)^2$
H	-0.7	0.00	0.00
2-Cl	-1.2	0.71	0.15
3-Cl	-0.7	0.71	0.12
4-Cl	-0.6	0.71	0.07
4-F	-1.0	0.14	0.01
4-NO <sub>2</sub>	-2.0	0.37	0.69
4-COOCH <sub>3</sub>	-1.7	-0.01	0.23
4-CH <sub>3</sub>	-1.2	0.56	0.03
2-OH	-0.4	-0.60	0.04
3-OH	-0.2	-0.60	0.00
4-OH	-0.1	-0.60	0.12
4-OCH <sub>3</sub>	-0.9	-0.06	0.06
2-NH <sub>2</sub>	-1.4	-1.23	0.32
4-N(CH <sub>3</sub> ) <sub>2</sub>	-2.0	0.18	0.67
3-Cl,4-Cl	-2.0	1.42	0.37
3-(CF <sub>3</sub> ) <sub>2</sub>	-1.1	1.96	0.77
2-Cl,6-Cl	-2.0	1.42	0.58
2-OH,6-OH	-0.7	-1.34	0.15
4-pyridyl	-0.9	-0.54	0.52

Table 14.2: Comparison of Experimental and Model Activities

	Substituents	$\pi$	$\sigma_M$	Model	$pI_{50}$ Experimental
1)	3-NHCO(CHOH) <sub>2</sub> H 5-NHCO(CHOH) <sub>2</sub> H	-3.812	0.064	-0.19	3.0
2)	3-NHCOCH <sub>2</sub> CH <sub>3</sub> 5-NHCOCH <sub>2</sub> CH <sub>3</sub>	-0.964	0.269	-0.68	2.5
3)	3-NHCOCH <sub>3</sub> 5-NHCOCH <sub>3</sub>	-1.552	0.391	-0.71	1.9
4)	3-NHCOCH <sub>3</sub> 5-OH	-1.763	0.301	-0.60	1.7
5)	3-NHCOCOOCH <sub>2</sub> CH <sub>3</sub> 5-NHCOCOOCH <sub>2</sub> CH <sub>3</sub>	-1.890	0.838	-1.40	1.7
6)	3-NHCOCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub> 5-NHCOCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	-0.376	0.147	-0.70	1.3
7)	3-NHCO(CHOH) <sub>2</sub> H	-1.906	0.032	-0.45	1.3
8)	3-NHCOCH <sub>3</sub> 5-NH <sub>2</sub>	-1.653	0.157	-0.49	1.0
9)	3-NHCOCH <sub>2</sub> CH <sub>3</sub>	-0.482	0.134	-0.68	0.7
10)	3-NHCOCH <sub>3</sub>	-0.776	0.196	-0.66	0.7

## 14.4 Target Transformation

Hansch[53] has tabulated  $\pi$  and  $\sigma$  values for many substituents.  $\sigma$  values are dependent upon the substituent's location in the ring. Tabular values are predominantly for para and meta locations.

I developed group contribution estimation techniques for  $\pi$ ,  $\sigma_M$ , and  $\sigma_P$ . The contributions were obtained by re-regressing the substituent constants tabulated by Hansch[53]. This procedure is similar to the one discussed in Chapter 5. However, my main objective was to obtain additive group contributions not group consistency.

$\pi$  values for 68 substituents were regressed yielding contributions for 24 groups. The average absolute error was 0.225. The average absolute value for  $\pi$  values is 0.738.  $\sigma_M$  values for 86 substituents were regressed yielding contributions for 29 groups. The average absolute error was 0.063. The average absolute  $\sigma_M$  value is 0.236.  $\sigma_P$  values for 93 substituents were regressed yielding contributions for 29 groups. The average absolute error was 0.121. The average absolute  $\sigma_P$  value is 0.275. The developed techniques are presented in Appendix A.

Overall the estimation techniques have considerable error. I use them only to demonstrate the interactive design's applicability to another type of target. The non-additive nature of  $\sigma_P$  and  $\sigma_M$  were previously noted[75]. Further investigation should concentrate on regressing original data to obtain additive group contributions and compare these to the accuracy obtained by the current group substituents.

Equation 14.6 shows that to perform an interactive design we would use a two dimensional  $\sigma$  vs.  $\pi$  physical property design space. Cramer et.al. used an average

Table 14.3: Consistent Groups for Drug Design

---

$-\text{CH}_3$	$-\text{CH}_2-$	$>\text{CH}-$	$>\text{C}<$	$=\text{CH}_2$	$=\text{CH}\cdot$
$=\text{C}<$	$-\text{F}$	$-\text{Cl}$	$-\text{Br}$	$-\text{I}$	$-\text{OH}$
$-\text{O}-$	$>\text{C}=\text{O}$	$-\text{CH}=\text{O}$	$-\text{COOH}$	$-\text{COO}-$	$-\text{NH}_2$
$-\text{NH}-$	$>\text{N}-$	$=\text{N}-$	$-\text{CN}$	$-\text{NO}_2$	$-\text{S}-$

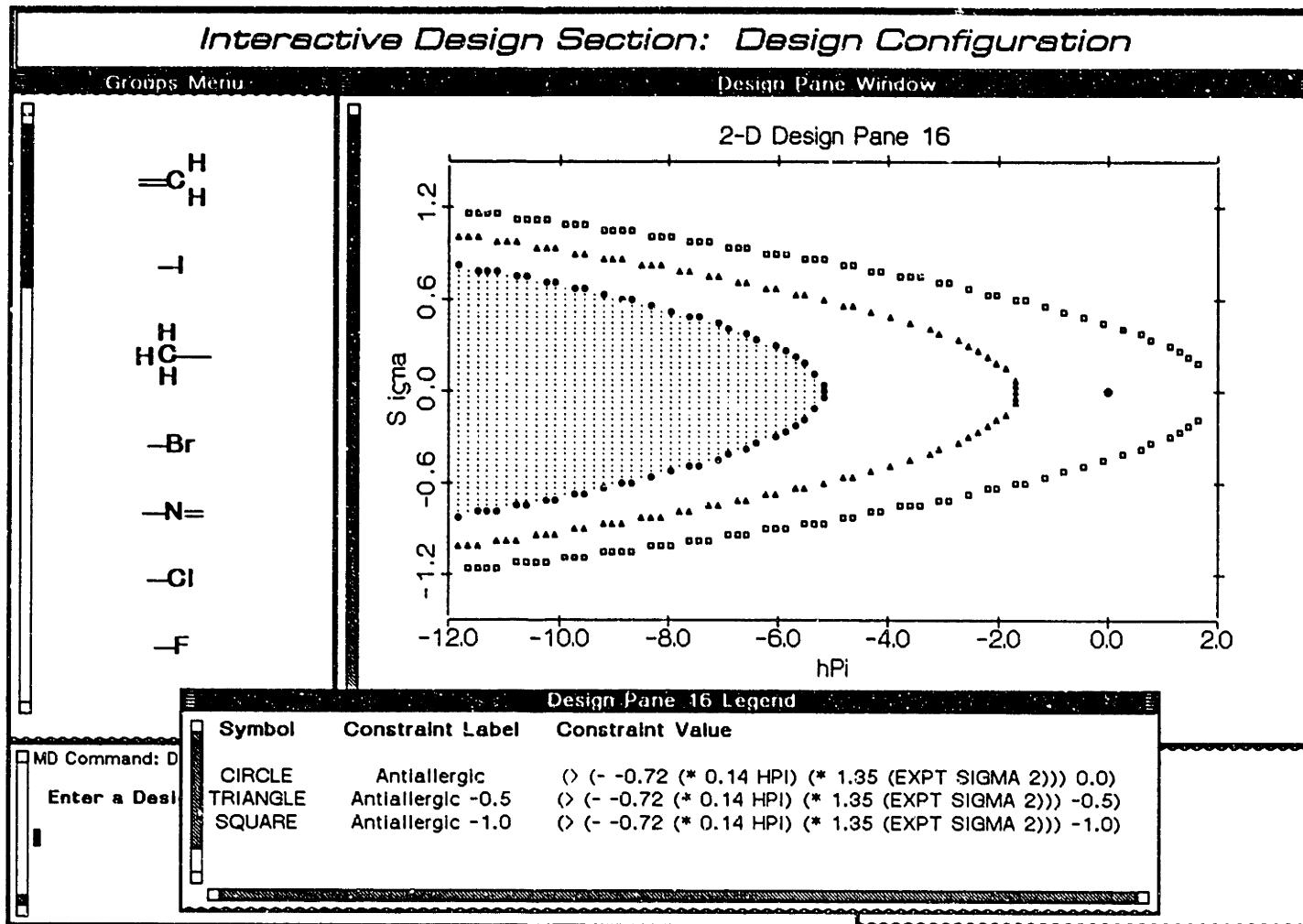
---

value for  $\sigma$ . I use only values for  $\sigma_M$ . This restricts me to design analogs with only meta substitutions. The consistent set of groups to be used in the design is shown in Table 14.3.

## 14.5 Interactive Design

Figure 14.3 shows our two dimensional  $\sigma_M$  vs.  $\pi$  design space. The contours represent solutions of Equation 14.6 with  $\text{p}I_{50}$  equal to  $-1.0$ ,  $-0.5$ , and  $0.0$ . The direction of maximum activity is thus near  $\sigma_M$  equal to zero and  $\pi$  large and negative. Figure 14.4 shows a pair of meta substituents in our target direction.

These meta substituents correspond to entry 1 of Table 14.2. This analog had a thousand times more activity over the original unsubstituted compound[22].

Figure 14.3:  $\sigma_M$  vs.  $\pi$  Drug Design Space

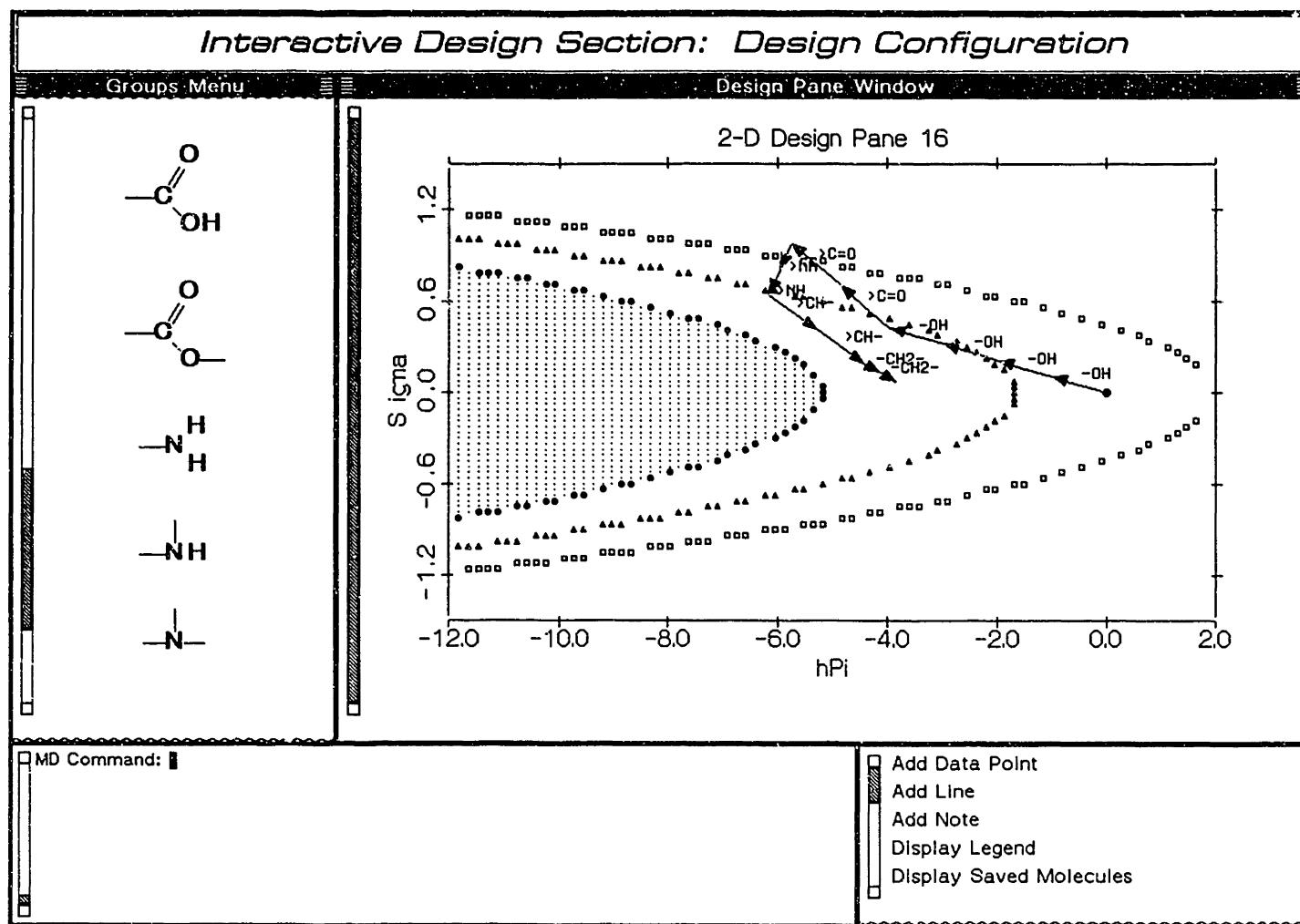


Figure 14.4: Two High Activity Meta-Substituents

# Chapter 15

## Conclusions

I have developed a systematic methodology for designing molecules possessing desired physical properties. Knowledge and information from a variety of fields were required in development. This requirement of diverse information is characteristic of many of the problems addressed in the field of artificial intelligence which attempt to capture “commonsense reasoning.” As more and more “real world” problems are addressed by the paradigms of AI more and more information will have to be systematized.

I have the following conclusions:

1. The design of molecules possessing desired physical properties is achievable by an efficient and informed methodology.
2. The generate and test paradigm is extremely useful for design problems. It provides a candidate solution which can be evaluated using a variety of knowledge sources.
3. Intervals are a powerful representation for design problems. Although I inves-

tigated interval arithmetic's representational power primarily from the context of molecular design, I believe it provides an excellent representation for many applications. Interval analysis has been commonly used in the field artificial intelligence[23,109]. It will become more commonly used in chemical engineering in the near future.

My methodology represents an initial attempt in systematizing those tasks needed in molecular design. As future research continues in the molecular design area I believe the generate and test paradigm and interval representation will continue to play an important role. I also believe that they will find usefulness in the more traditional chemical engineering design areas.

# Chapter 16

## Recommendations

I have several recommendations for improving and extending my methodology. These recommendations are in five areas:

- Add structural constraints. Search for a systematic way in which structural constraints can be checked for sufficiency.
- Investigate the performance of different meta-group division strategies.
- Investigate cooperation between interactive and automatic design methodologies.
- Development new, specific, design oriented estimation techniques.
- Investigate ways in which molecular structure can be displayed on the computer preferably in a two dimensional representation.

## 16.1 Structural Constraints

The structural constraints I used came from basic graph theoretic analysis of molecular structures. Additional structural constraints probably exist. Performing an automatic design and then examining the candidate molecules for structural infeasibility should be the first step in identifying new constraints. I believe more constraints involving mixed groups and bond types can be found.

The hope is that a theoretical approach to the molecule formation problem can be found. This would determine whether a given set of structural constraints is sufficient to generate structurally feasible molecules.

## 16.2 Meta-Group Strategies

After performing several automatic designs for a given problem one begins to identify which meta-group division strategies work better than others. Investigating this relative performance is an interesting issue. However, the specificity of the problem which must be addressed could possibly preclude the value of any finding. A meta-group strategy is dependent upon constraints, estimation procedures, meta-molecules, groups, and fundamental property selected.

## 16.3 Design Methodology Cooperation

The interactive and automatic design methodologies can cooperate in several ways. Two ways are:

1. Use automatic design to complete a molecule partially designed interactively. The interactive design could be initiated through an evolutionary desire. An existing molecule is brought into the design space. Non-desired groups are removed from the molecule. The automatic design searches for replacements to these removed groups.
2. Use the interactive design methodology to identify ways in which automatically designed molecules could be modified without effecting the desirability of their physical properties. The objective of the modification could be to create molecules more available or more easily synthesized.

## 16.4 Specific Estimation Techniques

Most estimation techniques are developed for applicability over a large range of compounds. To design refrigerants it is only necessary to have estimation techniques specific to C1-C5 compounds. Additionally, if one of our constraints requires the estimation of  $P_{vp}$  at 298K, then we should develop an estimation technique specific to refrigerants and specific to the temperature in question.

## 16.5 Molecular Display

The two dimensional display of molecular structure would benefit the evaluation of designed molecules. The major difficulty encountered in evaluating the enumeration and screening steps of the methodology was due to the inability to easily examine

molecular structure.

Three dimensional models are not as satisfactory as two dimensional for quick visualization of a molecular structure. However, three dimensional display is easier to do.

# Bibliography

- [1] Agoos, Alice: Serving up a better package for foods. *Chemical Week*, pp. 100–104, October 16, 1985.
- [2] Alternburg, von Kurt: Die Abhangigkeit der Siedetemperatur isomerer Kohlenwasserstoffe von der Form der Molekule. *Brennstoff Chemie*, pp. 331–336, **47**, 11, 1966.
- [3] American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.: *ASHRAE Handbook of Fundamentals*. ASHRAE, New York, 1972.
- [4] Asaithambi, N. S., Shen Zuhe, and R. E. Moore: On Computing the Range of Values. *Computing*, pp. 225–237, **28**, 1982.
- [5] Balaban, A. T.: *Chemical Applications of Graph Theory*. Academic Press, London, 1976.
- [6] Barton, Allan F. M.: *CRC Handbook of Solubility Parameters and Other Cohesive Parameters*. CRC Press Inc., Boca Raton, Florida, 1983.
- [7] Beck, Jacob: Effect of Orientation and of Shape Similarity on Perceptual Grouping. *Perception and Psychophysics*, pp. 300-302, **1**, 1966.
- [8] Berg, L.: Selecting the Agent for Distillation Processes. *Chemical Engineering Progress*, pp. 52–57, **65**, 9, September 1969.
- [9] Bobrow, Daniel G.: *Qualitative Reasoning about Physical Systems*. The MIT Press, Cambridge, Massachusetts, 1984.

- [10] Boff, Kenneth R., Lloyd Kaufman, and James P. Thomas, eds.: *Handbook of Perception and Human Performance. Volume 1. Sensory Processes and Perception.* John Wiley & Sons, New York, 1986.
- [11] Bondi, A.: van der Waals Volumes and Radii. *Journal of Physical Chemistry*, pp. 441-451, **68**, 3, March 1964.
- [12] Bondi, A.: Estimation of the Heat Capacity of Liquids. *Industrial and Engineering Chemistry. Fundamentals*, **5**, pp. 442–449, November 1966.
- [13] Brignole, Esteban A., Susana Bottini, and Rafiqul Gani: A Strategy for the Design and Selection of Solvents for Separation Processes. *Fluid Phase Equilibria*, pp. 125–132, **29**, 1986.
- [14] Broadbent, D. E.: *Perception and Communication*. Pergamon, New York, 1958.
- [15] Brown, W. V.: Economics of Recovering Acetic Acid. *Chemical Engineering Progress*, pp. 65–68, **59**, 10, October 1963.
- [16] Chen, Ning Hsing: Generalized Correlation for Latent Heat of Vaporization. *Journal of Chemical and Engineering Data*, pp. 207-210, **10**, 2, April 1965.
- [17] Chevier, J. C. and E. L. Yuan: Polyimide Dielectrics for High Density and High Speed Packaging. *Proceedings of the Sixth Annual International Electronics Packaging Conference*, pp. 285–293, 1986.
- [18] Clarke, Eric A.: Evaluation of the Molecular Design Approach in the Development of New Dyes. *Drug Metabolism Reviews*, pp. 997-1009, **15**, 5 & 6, 1984.
- [19] Cramer, Richard D., Kenneth M. Snader, Chester R. Willis, Lawrence W. Chakrin, Jean Thomas, and Blaine M. Sutton: Application of Quantitative Structure–Activity Relationships in the Development of the Antiallergic Pyranenamines. *Journal of Medicinal Chemistry*, pp. 714–725, **22**, 6, 1979.

- [20] Cramer, Richard D.: BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. *Journal of the American Chemical Society*, pp. 1837–1849, **102**, 6, March 12, 1980.
- [21] Cramer, Richard D.: BC(DEF) Parameters. 2. An Empirical Structure-Based Scheme for the Prediction of Some Physical Properties. *Journal of the American Chemical Society*, pp. 1849–1859, **102**, 6, March 12, 1980.
- [22] Cramer, Richard D.: A QSAR Success Story. *CHEMTECH*, pp. 744–747, December 1980.
- [23] Davis, Ernest: Constraint Propagation with Interval Labels. *Artificial Intelligence*, pp. 281–331, **32**, 1987.
- [24] Debye, P.: *Math. Vorlesungen Univ. Göttingen*; **6**, 19, 1914.
- [25] Del Re, Giuseppe: A Simple MO-LCAO Method for the Calculation of Charge Distributions in Saturated Organic Molecules. *Journal of the Chemical Society*, pp. 4031–4040, November 1958.
- [26] Derringer, George C. and Richard L. Markham: A Computer-Based Methodology for Matching Polymer Structures with Required Properties. *Journal of Applied Polymer Science*; **30**, pp. 4609–4617, 1985.
- [27] Dillinger, Thomas E.: *VLSI Engineering*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [28] Dossat, Roy J.: *Principles of Refrigeration*. John Wiley and Sons, New York, 1981.
- [29] Douglas, James M.: *Conceptual Design of Chemical Processes*. McGraw-Hill Book Company, New York, 1988.
- [30] DuBois, Grant E.: Nonnutritive Sweeteners. The Search for Sucrose Mimics. *Annual Reports in Medicinal Chemistry*, pp. 323–332, **17**, 1982.

- [31] Eaglesfield, P., B. K. Kelly, and J. F. Short: Recovery of Acetic Acid from Dilute Aqueous Solutions by Liquid-Liquid Extraction – Part 1. *The Industrial Chemist*, pp. 147–151, April 1953.
- [32] Eaglesfield, P., B. K. Kelly, and J. F. Short: Recovery of Acetic Acid from Dilute Aqueous Solutions by Liquid-Liquid Extraction – Part 2. *The Industrial Chemist*, pp. 147–151, June 1953.
- [33] Fedors, Robert F.: A Method for Estimating Both the Solubility Parameters and Molar Volumes of Liquids. *Polymer Engineering Science*, pp. 147–154, **14**, 2, February 1974.
- [34] Fedors, Robert F.: A Method for Estimating Both the Solubility Parameters and Molar Volumes of Liquids: Supplement. *Polymer Engineering Science*, p. 472, **14**, 6, June 1974.
- [35] Fedors, R. F.: A Relationship between Chemical Structure and the Critical Temperature. *Chemical Engineering Communications*, pp. 149–151, **16**, 1982.
- [36] Fenelon, Paul J.: Theoretical Prediction of Pressure Loss in Pressurized Plastic Containers. *Polymer Engineering and Science*, pp. 440–446, **13**, 6, November 1973.
- [37] Fogiel, Max: *Modern Microelectronics*. Research and Education Association, New York, 1972.
- [38] Francis, Alfred W.: Solvent Selectivity for Hydrocarbons. *Industrial and Engineering Chemistry*, pp. 764–771, **36**, 8, August 1944.
- [39] Franke, Rainer: *Theoretical Drug Design Methods*. Elsevier, Amsterdam, 1984.
- [40] Franklin, J. L.: Prediction of Heat and Free Energies of Organic Compounds. *Industrial and Engineering Chemistry*, pp. 1070–1076, **41**, May 1949.
- [41] Fredenslund, Aage, Jürgen Gmehling, and Peter Rasmussen: *Vapor-Liquid Equilibria using UNIFAC*. Elsevier, Amsterdam, 1977.

- [42] Gani, R. and E. A. Brignole: Molecular Design of Solvents for Liquid Extraction Based on UNIFAC. *Fluid Phase Equilibria*, **13**, pp. 331–340, 1983.
- [43] Gibson, J. J.: *The Perception of the Visual World*. Houghton Mifflin, Boston, 1950.
- [44] Godfrey, Norman B.: Solvent Selection via Miscibility Number. *CHEMTECH*, pp. 359–363, June, 1972.
- [45] Goosey, M. T.: Permeability of Coatings and Encapsulants for Electronic and Optoelectronic Devices. *Polymer Permeability*, J. Comyn, ed. Elsevier, London, pp. 309–339, 1985.
- [46] Gordon, M. and G. R. Scantlebury: Non-Random Polycondensation: Statistical Theory of the Substitution Effect. *Transactions of the Faraday Society*, pp. 604–621, **60**, 495, March 1964.
- [47] Gray, Neil A. B.: *Computer-Assisted Structure Elucidation*. John Wiley and Sons, New York, 1986.
- [48] Hagler, A. T.: Theoretical Simulation of Conformation, Energetics, and Dynamics of Peptides. *The Peptides*. Academic Press, London, **7**, pp. 213–299, 1985.
- [49] Hammett, Louis P.: Some Relations Between Reaction Rates and Equilibrium Constants. *Chemical Reviews*, **17**, 1, pp. 125–136, August 1935.
- [50] Hansch, Corwin, Robert M. Muir, Toshio Fujita, Peyton P. Maloney, Fred Geiger, and Margaret Streich: The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *Journal of the American Chemical Society*, **85**, pp. 2817–2824, September 10, 1963.
- [51] Hansch, Corwin and A. Ruth Steward: The Use of Substituent Constants in the Analysis of the Structure-Activity Relationship in Penicillin Derivatives. *Journal of Medicinal Chemistry*, pp. 691–694, **7**, 6, November 6, 1964.

- [52] Hansch, Corwin and Stefan H. Unger: Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *Journal of Medicinal Chemistry*, pp. 1217–1222, **16**, 11, 1973.
- [53] Hansch, Corwin and Albert Leo: *Substituent Constants for Correlation Analysis in Chemistry and Biology*. John Wiley and Sons, New York 1979.
- [54] Hayes-Roth, Frederick, Donald A. Waterman, and Douglas B. Lenat. *Building Expert Systems*. Addison-Wesley, Reading, Massachusetts, 1983.
- [55] Henry, Douglas R., Peter C. Jurs, and William A. Denny: Structure-Antitumor Activity Relationships of 9-Anilinoacridines Using Pattern Recognition. *Journal of Medicinal Chemistry*, pp. 899–908, **25**, 8, 1982.
- [56] Hildebrand, Joel H. and Robert L. Scott: *The Solubility of Nonelectrolytes*. Reinhold Publishing Company, New York, 1950.
- [57] Hildebrand, Joel H. and Robert L. Scott: *Regular Solutions*. Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
- [58] Himmelblau, D. M.: Material Balance Rectification via Interval Arithmetic. in *Process Systems Engineering*, The Institution of Chemical Engineers Symposium Series No. 92, pp. 121–132, 1985.
- [59] Hosoya, Haruo and Miyuki Murakami: Topological Index as Applied to  $\pi$ -Electronic Systems. II. Topological Bond Order. *Bulletin of the Chemical Society of Japan*, pp. 3512–3517, **48**, 12, 1975.
- [60] Hubel, D. H. and T. N. Wiesel: Brain Mechanisms of Vision. in *The Brain: A Scientific American Book*, W. H. Freeman, New York, 1979.
- [61] Joback, Kevin G.: A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques. Master's thesis, Massachusetts Institute of Technology, June 1984.

- [62] Joback, Kevin G. and Robert C. Reid: Estimation of Pure-Component Properties from Group-Contributions. *Chemical Engineering Communications*; **57**, pp. 233-243, 1987.
- [63] Johnson, Richard A. and Dean W. Wichern: *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1982.
- [64] Kaempf, G., H. Loewer, and M. W. Witman: Polymers as Substrates and Media for Data Storage. *Polymer Engineering and Science*, pp. 1421-1435, **27**, 19, October 1987.
- [65] Kier, Lemont B. and Lowell H. Hall: *Molecular Connectivity in Structure-Activity Analysis*. John Wiley & Sons, New York, 1986.
- [66] Kirk-Othmer Encyclopedia of Chemical Technology: *Barrier Polymers*. John Wiley and Sons, New York, 3rd Edition, Volume 3, 1978.
- [67] Klincewicz, Karen M.: *Prediction of Critical Temperatures, Pressures, and Volumes of Organic Compounds from Molecular Structure*. Master's thesis, Massachusetts Institute of Technology, June 1982.
- [68] Koros, W. J., B. J. Story, S. M. Jordan, K. O'Brien, and G. R. Husk: Material Selection Considerations for Gas Separation Processes. *Polymer Engineering and Science*, pp. 603-610, **27**, 8, April 1987.
- [69] Kosslyn, Stephen Michael: *Image and Mind*. Harvard University Press, Cambridge, Massachusetts, 1980.
- [70] Labuza, T. P., S. Mizrahi, and M. Karel: Mathematical Models for Optimization of Flexible Film Packaging of Foods for Storage. *Transactions of the ASAE*, pp. 150-155, 1972.
- [71] Langley, Billy C.: *Refrigeration and Air Conditioning*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

- [72] Lee, Byung Ik and Michael G. Kesler: A Generalized Thermodynamic Correlation Based on Three-Parameter Corresponding States. *AICHE J.*, pp. 510–527, **21**, 3, May 1975.
- [73] Lee, John F. and Francis Weston Sears: *Thermodynamics*. Addison-Wesley, Reading Massachusetts, 1963.
- [74] Lee W.M.: Selection of Barrier Materials from Molecular Structure. *Polymer Engineering and Science*, pp. 65–69, **20**, 1, Mid-January 1980.
- [75] Leo, Albert. Personal communication. January 1989.
- [76] Lindsay, Robert K., Bruce G. Buchanan, Edward A. Feigenbaum, Joshua Lederberg: *Applications of Artificial Intelligence for Organic Chemistry*. McGraw-Hill, New York, 1980.
- [77] Lo, Teh C., Malcolm H. I. Baird, and Carl Hanson: *Handbook of Solvent Extraction*. John Wiley & Sons, New York, 1983.
- [78] Lydersen, A. L.: Estimation of Critical Properties of Organic Compounds. University of Wisconsin College of Engineering, Engineering Experimental Station Report 3, Madison, Wisconsin, April 1955.
- [79] Lyman, Warren J., William F. Reehl, and David H. Rosenblatt: *Handbook of Chemical Property Estimation Methods*. McGraw-Hill Book Company, New York, 1982.
- [80] Magnussen, Thomas, Peter Rasmussen, and Aage Fredenslund: UNIFAC Parameter Table for Prediction of Liquid-Liquid Equilibria. *Industrial and Engineering Chemistry. Process Design and Development*, pp. 331–339, **20**, 1981.
- [81] Malinowski, Edmund R. and Darryl G. Howery: *Factor Analysis in Chemistry*. John Wiley and Sons, New York, 1980.
- [82] Mardia, K. V., J. T. Kent, and J. M. Biddy: *Multivariate Analysis*. Academic Press, London, 1979.

- [83] Mark, H. F.: Textile Science and Engineering: Present and Future. *Chemical Engineering Progress*, pp. 44–54, December 1987.
- [84] Martin, Yvonne Connolly: *Quantitative Drug Design: A Critical Introduction*. Marcel Dekker, Inc., New York, 1978.
- [85] Matson, S. L., W. J. Ward, S. G. Kimura, and W. R. Browall: Membrane Oxygen Enrichment II. *Journal of Membrane Science*, pp. 79–96, **29**, 1986.
- [86] McElroy, Michael, Ross J. Salawitch, Steven C. Wofsy, and Jennifer A. Logan: Reductions of Antarctic Ozone due to Synergistic Interactions of Chlorine and Bromine. *Nature*, pp. 759–762, **321**, 19, June 1986.
- [87] Mih, Winston C.: Catalysts for Epoxy Molding Compounds in Microelectronic Encapsulation. *Polymers in Electronics*, ACS Symposium Series 242, Theodore Davidson, ed. American Chemical Society, pp.273–283, 1984.
- [88] Miller, Donald G.: A Simple Reduced Equation for the Estimation of Vapor Pressures. *Journal of Physical Chemistry*, pp. 3209–3212, **69**, 9, September, 1965.
- [89] Moore, Ramon E.: *Methods and Applications of Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, 1979.
- [90] Nabors, Lyn O'Brien and Robert C. Gelardi: *Alternative Sweeteners*. Marcel Dekker, Inc., New York, 1986.
- [91] Nelson, R. C., V. F. Figurelli, J. G. Walsham, and G. D. Edwards: Solution Theory and the Computer – Effective Tools for the Coatings Chemist. *Journal of Paint Technology*, pp. 644–652, **42**, 550, November 1970.
- [92] Nirmalakhandan, Nagamany and Richard E. Speece: Structure-Activity Relationships. *Environmental Science & Technology*, pp. 606–615, **22**, 6, 1988.
- [93] Pensak, David A.: Modeling System Cuts Wasteful Steps from Complex R&D. *Industrial Research & Development*, pp. 74–78, January 1983.

- [94] Perry, Robert H. and Cecil H. Chilton: *Chemical Engineers' Handbook*. McGraw-Hill Book Company, New York, 1973.
- [95] Rall, L.B.: Improved Interval Bounds for Ranges of Functions. in Lecture Notes in Computer Science 212: *Interval Mathematics 1985*, Proceedings of the International Symposium, K. Nickel, ed., Springer-Verlag, Berlin, 1985.
- [96] Randić, Milan: On Characterization of Molecular Branching. *Journal of the American Chemical Society*, pp. 6609-6615, **97**, 23, November 12, 1975.
- [97] Rao, R.: *J. Chem. Phys.*, **9**, 682, 1941.
- [98] Ratschek, H. and J. Rokne: *Computer Methods for the Range of Functions*. John Wiley & Sons, New York, 1984.
- [99] Ray, Louis C. and Russell A. Kirsch: Finding Chemical Records by Digital Computers. *Science*, pp. 814-819, **126**, October 25, 1957.
- [100] Reed, Thomas M. and Keith E. Gubbins: *Applied Statistical Mechanics*. McGraw-Hill Book Company, New York, 1973.
- [101] Reid, R. C., J. M. Prausnitz, and T. K. Sherwood: *The Properties of Gases and Liquids*. McGraw-Hill Book Company, New York, 1977.
- [102] Reid, R. C., J. M. Prausnitz, and B. E. Poling: *The Properties of Gases and Liquids*. McGraw-Hill Book Company, New York, 1987.
- [103] Rheineck, A. E. and K. F. Lin: Solubility Parameter Calculations Based on Group Contributions. *Journal of Paint Technology*, pp. 611-616, **40**, 527, December 1968.
- [104] Riedel, L.: Kritischer Koeffizient, Dichte des gesättigen Dampfes und Verdampfungswärme. *Chemie Ingenieur Technik*, pp. 679-683, **26**, 12, 1954.
- [105] Riordan, John: *An Introduction to Combinatorial Analysis*. John Wiley & Sons, New York, 1958.

- [106] Ris, Frederic N.: Tools for the Analysis of Interval Arithmetic. pp. 75–98. in *Lecture Notes in Computer Science 29: Interval Mathematics 1975*, Proceedings of the International Symposium, K. Nickel, ed., Springer-Verlag, Berlin, 1975.
- [107] Rouvray, Dennis H.: Predicting Chemistry from Topology. *Scientific American*, pp. 40–47, **255**, 3, September 1986.
- [108] Rowlinson, J. S.: *Liquids and Liquid Mixtures*. Butterworth, London, 1969.
- [109] Sacks, Elisha P.: *Hierarchical Inequality Reasoning*. Massachusetts Institute of Technology, Laboratory for Computer Science, Technical Memo 312, February 1987.
- [110] Salame, Morris: Prediction of Gas Barrier Properties of High Polymers. *Polymer Engineering and Science*, pp. 1543–1546, **26**, 33, December 1986.
- [111] Satoh, S.: *Journal Sci. Research Inst.*, **43**, 79, 1948.
- [112] Schuyer, J.: Sound Velocity in Polyethylene. *J. Polymer Sci.*, **36**, pp. 475–483, 1959.
- [113] Sibley, Howard W.: Selecting Refrigerants for Process Systems. *Chemical Engineering*, pp. 71–76, May 16, 1983.
- [114] Simon, H. A.: What is visual imagery? An information processing interpretation. in *Cognition in Learning and Memory*, L. W. Gregg, ed., John Wiley & Sons, New York, 1972.
- [115] Small, P. A.: Some Factors Affecting the Solubility of Polymers. *Journal of Applied Chemistry*, pp. 71–80, **3**, February 1953.
- [116] Solomon, Susan, Rolando R. Garcia, F. Sherwood Rowland, and Donald J. Wuebbles: On the Depletion of Antarctic Ozone. *Nature*, pp. 755–758, **321**, 19, June 1986.

- [117] Steingiser, S., S. P. Nemphos, and M. Salame: Barrier Polymers. Kirk-Othmer Encyclopedia of Chemical Technology, 3<sup>rd</sup> edition, Volume 3. John Wiley & Sons, New York, 1978.
- [118] Stephanopoulos, George and D. W. Townsend: Synthesis in Process Development. *Chemical Engineering Research and Development*, pp. 160–174, 64, May 1986.
- [119] Stillings, Neil A., Mark H. Feinstein, Jay L. Garfield, Edwina L. Rissland, David A. Rosenbaum, Steven E. Weisler, and Lynne Baker-Ward: *Cognitive Science: An Introduction*. The MIT Press, Cambridge, Massachusetts, 1987.
- [120] Stuper, Andrew J., William E. Brügger, and Peter C. Jurs: *Computer Assisted Studies of Chemical Structure and Biological Function*. John Wiley & Sons, New York, 1979.
- [121] Sussenguth, E. H.: A Graph-Theoretic Algorithm for Matching Chemical Structures. *Journal of Chemical Documentation*, pp. 36–43, 5, 1965.
- [122] Taft, R. W.: Separation of Polar, Steric, and Resonance Effects in Reactivity. in *Steric Effects in Organic Chemistry*. M. S. Newman, ed. John Wiley & Sons, New York, 1956.
- [123] Thieler, P.: Technical Calculations by Means of Interval Mathematics. in Lecture Notes in Computer Science 212: *Interval Mathematics 1985*, Proceedings of the International Symposium, K. Nickel, ed., Springer-Verlag, Berlin, 1985.
- [124] Tortorello, Anthony and Mary A. Kinsella: Solubility Parameter Concept in the Design of Polymers for High Performance Coatings I. *Journal of Coatings Technology*, pp. 99–38, 55, 696, January 1983.
- [125] Tortorello, Anthony and Mary A. Kinsella: Solubility Parameter Concept in the Design of Polymers for High Performance Coatings II. *Journal of Coatings Technology*, pp. 29–38, 55, 697, February 1983.

- [126] Treybal, Robert E.: *Liquid Extraction*. McGraw-Hill Book Company, New York, 1963.
- [127] van Krevelen, D. W.: *Properties of Polymers: Correlations with Chemical Structure*. Elsevier, Amsterdam, 1972.
- [128] van Krevelen, D. W. and P. J. Hoflyzer. Unpublished. Cited in: van Krevelen, D. W.: *Properties of Polymers: Correlations with Chemical Structure*. Elsevier, Amsterdam, 1972.
- [129] Verloop, A.: The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies. in *Drug Design*, E. J. Ariens, ed. Academic Press, New York, 1972.
- [130] Vetere A.: New Generalized Correlations for Enthalpy of Vaporization of Pure Compounds, Laboratori Ricerche Chimica Industriale, SNAM PROGETTI, San Donato Milanese, 1973.
- [131] Watson, K. M.: Thermodynamics of the Liquid State. *Industrial and Engineering Chemistry*, pp. 398-406, **35**, April, 1943.
- [132] Wiener, Harry: Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *Journal of the American Chemical Society*, pp. 2636-2638, **69**, 11, November, 1947.
- [133] Winston, Patrick H.: *Artificial Intelligence*. Addison-Wesley, Reading, Massachusetts, 1984.
- [134] World Meteorological Organization Global Ozone Research and Monitoring Project Report No. 16: *Atmospheric Ozone 1985*. National Aeronautics and Space Administration, 1985.

# Appendix A

## Estimation Techniques

Most design procedures are based on models. Group contribution and equation oriented estimation techniques are the models of my molecular design procedure. In this appendix I present the physical property estimation techniques I used in my thesis work. Table A.1 lists the physical properties estimated by the techniques in this appendix.

### A.1 Normal Boiling Point

The normal boiling point,  $T_b$ , for a pure compound is the temperature at which the compound's vapor pressure equals 1 atmosphere. Two estimation techniques were used to estimate  $T_b$ . The first is a group contribution estimation technique the second is an equation oriented estimation technique.

Table A.1: Estimated Physical Properties

Section		Properties
A.1	$T_b$	Normal Boiling Point
A.5	$P_{vp}$	Vapor Pressure
A.7	$\Delta H_{vb}$	Enthalpy of Vaporization at $T_b$
A.8	$\Delta H_v$	Enthalpy of Vaporization
A.11	$C_{pL}$	Liquid Heat Capacity
A.6	$\omega$	Acentric Factor
A.3	$T_c$	Critical Temperature
A.4	$P_c$	Critical Pressure
A.2	$T_{br}$	Reduced Boiling Point
A.10	$C_{pv}^o$	Ideal Heat Capacity
A.9	$F_1$	Factor 1
A.9	$F_2$	Factor 2
A.9	$F_3$	Factor 3
A.12	$T_g$	Glass Transition Temperature
A.13	$\pi$	Permachor
A.14	$R$	Volume Resistivity
A.15	$P_{LL}$	Molar Polarization
A.16	$V$	Molar Volume
A.17	$M_w$	Molecular Weight
A.18	$\lambda$	Thermal Conductivity
A.19	$C_{ps}$	Solid Heat Capacity
A.20	$U$	Rao Function
A.21	$\delta_p$	Solubility Parameter Polar Component
A.21	$\delta_h$	Solubility Parameter Hydrogen Component

### A.1.1 Group Contribution Technique

Joback[62] developed a group contribution estimation technique for  $T_b$ . The model for the technique is given by:

$$T_b(K) = 198.18 + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,T_b} \quad (\text{A.1})$$

The values for the group contributions are given in Table A.2.

**Example:** Estimation of the normal boiling point of tripropyl amine:

Group	Occurrences	c Contribution	Total
-CH <sub>3</sub>	3	23.58	70.74
-CH <sub>2</sub> -	6	22.88	137.28
>N-	1	11.74	11.74
Total Contributions: 219.76			

$$T_b = 198.18 + 219.76 = 417.94K. \quad (\text{A.2})$$

The literature value is reported as 428K[61]. Joback[61] reported average estimation errors of 15K.

### A.1.2 Equation Oriented Technique

In a factor analytic study of physical properties  $T_b$  was found to be related to three factors[61]:

$$T_b = 358.39 - 25.26F_1 + 1.20F_2 - 64.94F_3 \quad (\text{A.3})$$

Table A.2:  $T_b$  Group Contributions

Groups	Contribution
Acyclic Increments	
$-\text{CH}_3$	23.58
$-\text{CH}_2-$	22.88
$>\text{CH}-$	21.74
$>\text{C}<$	18.25
$=\text{CH}_2$	18.18
$=\text{CH}-$	24.96
$=\text{C}<$	24.14
$=\text{C}=$	26.15
$\equiv\text{CH}$	9.20
$\equiv\text{C}-$	27.38
Ring Increments	
$>\text{CH}_2$	27.15
$>\text{CH}-$	21.78
$>\text{C}<$	21.32
$=\text{CH}-$	26.73
$=\text{C}<$	31.01
Halogen Increments	
$-\text{F}$	-0.03
$-\text{Cl}$	38.13
$-\text{Br}$	66.86
$-\text{I}$	93.84
Oxygen Increments	
$-\text{OH}$ (alcohol)	92.88
$-\text{OH}$ (phenol)	76.34
$-\text{O}-$ (nonring)	22.42
$-\text{O}-$ (ring)	31.22
$>\text{CO}$ (nonring)	76.75
$>\text{CO}$ (ring)	94.97
$-\text{CHO}$ (aldehyde)	72.24
$-\text{COOH}$ (acid)	169.09
$-\text{COO}-$ (ester)	81.10
$=\text{O}$ (except for above)	-10.50

Table A.2 Continued:  $T_b$  Group Contributions

Groups	Contribution
Nitrogen Increments	
-NH <sub>2</sub>	72.23
>NH (nonring)	50.17
>NH (ring)	52.82
>N- (nonring)	11.74
=N- (nonring)	74.60
=N- (ring)	83.08
-CN	125.66
-NO <sub>2</sub>	152.54
Sulfur Increments	
-SH	63.56
-S- (nonring)	68.78
-S- (ring)	52.10

Each of the three factors are estimated by group contribution techniques described in Section A.9.

## A.2 Reduced Boiling Point

The reduced boiling point for a pure compound is defined by:

$$T_{br} = \frac{T_b}{T_c} \quad (A.4)$$

Estimating  $T_{br}$  enables one to estimate  $T_c$  provided  $T_b$  is known. Additionally  $T_{br}$  is widely used in other equation oriented estimation techniques. A group contribution estimation technique was used in my thesis to estimate  $T_{br}$ .

### A.2.1 Group Contribution Technique

Joback[62] developed a group contribution estimation technique for  $T_{br}$ . The model for the technique, a modification of Lydersen's[78], is given by:

$$T_{br} = 0.584 + 0.965 \sum_{\text{all groups}} n_i \Delta_{i,T_{br}} - \left( \sum_{\text{all groups}} n_i \Delta_{i,T_{br}} \right)^2 \quad (\text{A.5})$$

The values for the group contributions are given in Table A.3.

**Example:** Estimation of the reduced boiling point of chlorotrifluoroethene:

Group	Occurrences	Contribution	Total
=C<	2	$1.17 \times 10^{-2}$	$2.34 \times 10^{-2}$
-Cl	1	$1.05 \times 10^{-2}$	$1.05 \times 10^{-2}$
-F	3	$1.11 \times 10^{-2}$	$3.33 \times 10^{-2}$
Total Contributions: $6.72 \times 10^{-2}$			

$$T_{br} = 0.584 + 0.965(6.72 \times 10^{-2}) - (6.72 \times 10^{-2})^2 = 0.644. \quad (\text{A.6})$$

The literature value is 0.647[102].

### A.3 Critical Temperature

Using the definition of  $T_{br}$ :

$$T_{br} = \frac{T_b}{T_c} \quad (\text{A.7})$$

we can estimate  $T_c$  providing we know values for  $T_{br}$  and  $T_b$ . Additionally I used an equation oriented technique.

Table A.3:  $T_{br}$  Group Contributions

Groups	Contribution $\times 10^2$
Acyclic Increments	
$-\text{CH}_3$	1.41
$>\text{CH}_2$	1.89
$>\text{CH}-$	1.64
$>\text{C}<$	0.67
$=\text{CH}_2$	1.13
$=\text{CH}-$	1.29
$=\text{C}<$	1.17
$=\text{C}=$	0.26
$\equiv\text{CH}$	0.27
$\equiv\text{C}-$	0.20
Ring Increments	
$>\text{CH}_2$	1.00
$>\text{CH}-$	1.22
$>\text{C}<$	0.42
$=\text{CH}-$	0.82
$=\text{C}<$	1.43
Halogen Increments	
$-\text{F}$	1.11
$-\text{Cl}$	1.05
$-\text{Br}$	1.33
$-\text{I}$	0.68
Oxygen Increments	
$-\text{OH}$ (alcohol)	7.41
$-\text{OH}$ (phenol)	2.40
$-\text{O}-$ (nonring)	1.68
$-\text{O}-$ (ring)	0.98
$>\text{CO}$ (nonring)	3.80
$>\text{CO}$ (ring)	2.84
$-\text{CHO}$ (aldehyde)	3.79
$-\text{COOH}$ (acid)	7.91
$-\text{COO}-$ (ester)	4.81
$=\text{O}$ (except for above)	1.43

Table A.3 Continued:  $T_{br}$  Group Contributions

Groups	Contribution $\times 10^2$
Nitrogen Increments	
-NH <sub>2</sub>	2.43
>NH (nonring)	2.95
>NH (ring)	1.30
>N- (nonring)	1.69
=N- (nonring)	2.55
=N- (ring)	0.85
-CN	4.96
-NO <sub>2</sub>	4.37
Sulfur Increments	
-SH	0.31
-S- (nonring)	1.19
-S- (ring)	0.19

### A.3.1 Equation Oriented Technique

In a factor analytic study it was found that  $T_c$  was related to three factors[61]:

$$T_c = 545.96 - 24.65F_1 + 11.99F_2 - 87.92F_3. \quad (A.8)$$

Each of the three factors are estimated by group contribution techniques described in Section A.9.

## A.4 Critical Pressure

The critical pressure of a pure compound is defined as the vapor pressure of the compound at its critical temperature. Two techniques were used to estimate  $P_c$ . The first is a group contribution estimation technique and the second is an equation oriented estimation technique.

### A.4.1 Group Contribution Technique

Joback[62] developed a group contribution estimation technique for  $P_c$ . The model for the technique is given by:

$$1/\sqrt{P_c} = 0.113 + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,P_c} \quad (\text{A.9})$$

$P_c$  has units of bar. The values for the group contributions are given in Table A.4.

**Example:** Estimation of the critical pressure of isobutyl acetate:

Group	Occurrences	Contribution	Total
$-\text{CH}_3$	4	$14.0 \times 10^{-3}$	$56.0 \times 10^{-3}$
$>\text{C}<$	1	$-1.1 \times 10^{-3}$	$-1.1 \times 10^{-3}$
$-\text{COO}-$	1	$9.1 \times 10^{-3}$	$9.1 \times 10^{-3}$
Total Contributions:			$6.4 \times 10^{-3}$

$$P_c = (0.113 + 6.4 \times 10^{-3})^{-2} = 31.9 \quad (\text{A.10})$$

The literature value is 30.2 bar[102]. Joback[61] reported an average error of 2.1 bar.

### A.4.2 Equation Oriented Technique

In a factor analytic study of physical properties  $P_c$  was found to be related to three factors[61]:

$$P_c = (0.1572 - 0.0193F_1 + 0.000678F_2 - 0.000330F_3)^{-2} \quad (\text{A.11})$$

Each of the three factors are estimated by group contribution techniques described in Section A.9.

Table A.4:  $P_c$  Group Contributions

Groups	Contribution
Acyclic Increments	
$-\text{CH}_3$	0.014
$>\text{CH}_2$	0.0096
$>\text{CH}-$	0.0044
$>\text{C}<$	-0.0011
$=\text{CH}_2$	0.0124
$=\text{CH}-$	0.007
$=\text{C}<$	0.0021
$=\text{C}=$	0.0004
$\equiv\text{CH}$	0.0072
$\equiv\text{C}-$	0.0016
Ring Increments	
$>\text{CH}_2$	0.0071
$>\text{CH}-$	0.006
$>\text{C}<$	-0.0029
$=\text{CH}-$	0.0053
$=\text{C}<$	0.0024
Halogen Increments	
$-\text{F}$	0.0089
$-\text{Cl}$	0.0081
$-\text{Br}$	-0.0025
$-\text{I}$	0.0066
Oxygen Increments	
$-\text{OH}$ (alcohol)	-0.0048
$-\text{OH}$ (phenol)	-0.012
$-\text{O}-$ (nonring)	0.0017
$-\text{O}-$ (ring)	-0.0016
$>\text{CO}$ (nonring)	0.0033
$>\text{CO}$ (ring)	0.0036
$-\text{CHO}$ (aldehyde)	0.0066
$-\text{COOH}$ (acid)	0.0051
$-\text{COO}-$ (ester)	0.0091
$=\text{O}$ (except for above)	-0.0069

Table A.4 Continued:  $P_c$  Group Contributions

Groups	Contribution
Nitrogen Increments	
-NH <sub>2</sub>	-0.0045
>NH (nonring)	-0.0013
>NH (ring)	-0.005
>N- (nonring)	-0.0042
=N- (nonring)	0.0131
=N- (ring)	-0.0044
-CN	0.0165
-NO <sub>2</sub>	0.0032
Sulfur Increments	
-SH	-0.002
-S- (nonring)	-0.0017
-S- (ring)	-0.0019

## A.5 Vapor Pressure

The pressure exerted by the vapor phase of a pure liquid in equilibrium with its liquid phase is defined as the vapor pressure of the compound. At equilibrium the equality of chemical potential, temperature, and pressure in both phases leads to the Clausius-Clapeyron equation:

$$\frac{d P_{vp}}{dT} = \frac{\Delta H_v}{T \Delta V_v}. \quad (\text{A.12})$$

Most vapor pressure estimation and correlation equations stem from an integration of Equation A.12. An equation oriented estimation technique was used in my thesis.

### A.5.1 Equation Oriented Technique

In my thesis I estimated vapor pressure using the Riedel-Plank-Miller equation oriented estimation technique. The equation oriented technique models vapor pressure as a function of the critical pressure, reduced boiling point, and the normal boiling point. This model is given by the following four equations:

$$\ln P_{vp_r} = -\frac{G}{T_r} [1 - T_r^2 + k (3 + T_r) (1 - T_r)^3] \quad (A.13)$$

$$G = 0.4835 + 0.4605 h \quad (A.14)$$

$$k = \frac{h/G - (1 + T_{b_r})}{(3 + T_{b_r}) (1 - T_{b_r})^2} \quad (A.15)$$

$$h = T_{b_r} \frac{\ln(P_c/1.01325)}{1 - T_{b_r}} \quad (A.16)$$

$T_b$  and  $T_c$  have units of K.  $P_c$  has units of bar.

**Example:** Estimating the vapor pressure of isopropanol at 450K:

Values for the required properties are:  $T_b = 355.4$  K,  $T_c = 508.3$  K, and  $P_c = 47.6$  bar.

$$T_r = 0.885$$

$$T_{b_r} = \frac{355.4}{508.3} = 0.699$$

$$h = 0.699 \frac{\ln(47.6/1.01325)}{1 - 0.699} = 8.94$$

$$g = 0.4835 + 0.4605(8.94) = 4.60$$

$$k = \frac{8.94/4.60 - (1 + 0.699)}{(3 + 0.699)(1 - 0.699)^2} = 0.729$$

The estimated vapor pressure at 450K is 15.09 bar. The literature value is 16.16 bar[102].

## A.6 Acentric Factor

The acentric factor was originally proposed to represent the acentricity or nonsphericity of a molecule. For monotonic gases,  $\omega$  is essentially zero. For methane it is still very small. However, for higher molecular weight hydrocarbons,  $\omega$  increases and often rises with polarity[101].

The acentric factor is defined by:

$$\omega = -(\log P_{vp_r})_{at\ T_r=0.7} - 1.000 \quad (A.17)$$

In my thesis  $\omega$  was estimated using an equation oriented estimation technique.

### A.6.1 Equation Oriented Technique

In my thesis I estimated the acentric factor using the Lee and Kesler's equation oriented estimation technique. The equation oriented technique models the acentric factor as a function of the critical pressure and reduced boiling point. This model is given by the following equation:

$$\omega = \frac{-\ln P_c - 5.92714 + 6.09648/T_{br} + 1.28862 \ln T_{br} - 0.169347 T_{br}^6}{15.2518 - 15.6875/T_{br} - 13.4721 \ln T_{br} + 0.43577 T_{br}^6} \quad (A.18)$$

**Example:** Estimating the acentric factor for perfluorohexane:

Values for the required properties are:  $T_b = 329.8$  K,  $T_c = 448.8$  K, and  $P_c = 18.7$  bar.

$$T_{b_r} = \frac{329.8}{448.8} = 0.735$$

The estimated value for  $\omega$  is 0.525. The literature value is 0.514[101].

## A.7 Enthalpy of Vaporization at $T_b$

The enthalpy of vaporization is the difference between the enthalpy of the saturated vapor and the enthalpy of the saturated liquid at the same temperature. Two techniques for estimating  $\Delta H_{vb}$  were used in my thesis. The first is a group contribution estimation technique the second is an equation oriented estimation technique.

### A.7.1 Group Contribution Technique

Joback[62] developed a group contribution estimation technique for  $\Delta H_{vb}$ . The model for the technique is given by:

$$\Delta H_{vb} = 15.30 + \sum_{\text{all groups}} n_i \Delta_{i,\Delta H_{vb}}. \quad (\text{A.19})$$

$\Delta H_{vb}$  has units of kJ/g-mol. The values for the group contributions is given in Table A.5.

**Example:** Estimating the enthalpy of vaporization at  $T_b$  of isobutyric acid:

Table A.5:  $\Delta H_{vb}$  Group Contributions

Groups	Contribution
Acyclic Increments	
-CH <sub>3</sub>	2.373
>CH <sub>2</sub>	2.226
>CH-	1.691
>C<	0.636
=CH <sub>2</sub>	1.724
=CH-	2.205
=C<	2.138
=C=	2.661
≡CH	1.155
≡C-	3.302
Ring Increments	
>CH <sub>2</sub>	2.398
>CH-	1.942
>C<	0.644
=CH-	2.544
=C<	3.059
Halogen Increments	
-F	-0.670
-Cl	4.532
-Br	6.582
-I	9.520
Oxygen Increments	
-OH (alcohol)	16.826
-OH (phenol)	12.499
-O- (nonring)	2.410
-O- (ring)	4.682
>CO (nonring)	8.972
>CO (ring)	6.645
-CHO (aldehyde)	9.093
-COOH (acid)	19.537
-COO- (ester)	9.633
=O (except for above)	5.909

Table A.5 Continued:  $\Delta H_{vb}$  Group Contributions

Groups	Contribution
Nitrogen Increments	
-NH <sub>2</sub>	10.788
>NH (nonring)	6.436
>NH (ring)	6.930
>N- (nonring)	1.896
=N- (nonring)	3.335
=N- (ring)	6.528
=NII	12.169
-CN	12.851
-NO <sub>2</sub>	16.738
Sulfur Increments	
-SH	6.884
-S- (nonring)	6.817
-S- (ring)	5.984

Group	Occurrences	Contribution	Total
-CH <sub>3</sub>	2	2.373	4.746
>CH-	1	1.691	1.691
-COOH	1	19.537	19.537

Total Contributions: 25.974

$$\Delta H_{vb} \text{ (kJ/mol)} = 15.30 + 25.974 = 41.27 \quad (\text{A.20})$$

The literature value is reported as 41.12 kJ/mol[101]. Joback[61] reported average estimation errors of 2.1%.

## A.7.2 Equation Oriented Technique

In a factor analytic study of physical properties  $\Delta H_{vb}$  was found to be related to three factors[61]:

$$\Delta H_{vb} \text{ (kJ/g - mol)} = 32.171 - 1.810F_1 - 1.068F_2 - 6.757F_3 \quad (\text{A.21})$$

Each of the three factors are estimated by group contribution techniques described in Section A.9.

## A.8 Enthalpy of Vaporization

The enthalpy of vaporization decreases with temperature and is zero at the critical point. Section A.7 discussed the techniques I use for estimating  $\Delta H_{vb}$ . To estimate  $\Delta H_v$  at temperatures other than the normal boiling point I use an equation oriented technique.

### A.8.1 Equation Oriented Technique

The Watson relation[131] is a widely used correlation between  $\Delta H_v$  and temperature. Knowing the enthalpy of vaporization at one temperature, the Watson relation enables you to compute the enthalpy of vaporization at another temperature knowing only the compound's critical temperature. I use  $\Delta H_{vb}$  as the original enthalpy of vaporization. The estimation model then becomes:

$$\Delta H_v = \Delta H_{vb} \left( \frac{T_c - T}{T_c - T_b} \right)^{0.38} \quad (\text{A.22})$$

## A.9 Factors

The statistical technique of factor analysis and the various applications of it involving physical properties are described in Appendix E. In one factor analytic study Joback[61] found that the nine physical properties:

$$\begin{array}{lll} V_c & 1/\sqrt{P_c} & C_{p,298}^o \\ n_A & \Delta G_{f,298}^o & \Delta H_{f,298}^o \\ T_b & T_c & \Delta H_{vb} \end{array}$$

could all be linearly related to three factors. Group contribution estimation techniques were developed for each of these three factors.

### A.9.1 $F_1$ Group Contribution Technique

Joback[61] proposed a group contribution technique for estimating  $F_1$ . I modified this technique to include some additional groups. The model for the technique is given by:

$$F_1 = -1.4545 + \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,F_1}. \quad (\text{A.23})$$

The values for the group contributions are given in Table A.6.

**Example:** Estimating  $F_1$  for 2,2-dimethylpentane:

Group	Occurrences	Contribution	Total
$-\text{CH}_3$	4	1.240	4.96
$-\text{CH}_2-$	2	-0.537	-1.074
$>\text{C}<$	1	-4.410	-4.410
Total Contributions: -0.524			

Table A.6:  $F_1$  Group Contributions

Groups	Contribution
<b>Acyclic Increments</b>	
$-\text{CH}_3$	1.240
$>\text{CH}_2$	-40.537
$>\text{CH}-$	-2.394
$>\text{C}<$	-4.410
$=\text{CH}_2$	1.307
$=\text{CH}-$	-0.350
$=\text{C}<$	-2.100
$=\text{C}=$	-0.206
$\equiv\text{CH}$	1.394
$\equiv\text{C}-$	-0.137
<b>Halogen Increments</b>	
$-\text{F}$	1.502
$-\text{Cl}$	1.692
$-\text{Br}$	1.903
$-\text{I}$	1.933
<b>Oxygen Increments</b>	
$-\text{OH}$ (alcohol)	2.173
$-\text{O}-$ (nonring)	-0.259
$>\text{CO}$ (nonring)	-0.208
$-\text{CHO}$ (aldehyde)	1.446
$-\text{COOH}$ (acid)	1.278
$=\text{O}$ (except for above)	1.832
<b>Nitrogen Increments</b>	
$-\text{NH}_2$	1.202
$>\text{NH}$ (nonring)	-0.644
$>\text{N}-$ (nonring)	-2.521
$-\text{CN}$	1.485
$-\text{NO}_2$	1.224
<b>Sulfur Increments</b>	
$-\text{SH}$	-0.044
$-\text{S}-$ (nonring)	1.455

$$F_1 = -1.4545 - 0.524 = -1.9785. \quad (\text{A.24})$$

The literature value is reported as -1.7662[61].

### A.9.2 $F_2$ Group Contribution Technique

Joback[61] proposed a group contribution technique for estimating  $F_2$ . I modified this technique to include some additional groups. The model for the technique is given by:

$$F_2 = -2.000 + \sum_{\text{all groups}} n_i \Delta_{i,F_2}. \quad (\text{A.25})$$

The values for the group contributions are given in Table A.7.

**Example:** Estimating  $F_2$  for 2,2-dimethylpentane:

Group	Occurrences	Contribution	Total
-CH <sub>3</sub>	4	0.974	3.896
-CH <sub>2</sub> -	2	-0.057	-0.114
>C<	1	-2.111	-2.111
Total Contributions:			1.671

$$F_2 = -2.000 + 1.671 = -0.329. \quad (\text{A.26})$$

The literature value is reported as -0.420[61].

Table A.7:  $F_2$  Group Contributions

Groups	Contribution
<b>Acyclic Increments</b>	
$-\text{CH}_3$	0.974
$>\text{CH}_2$	-0.057
$>\text{CH}-$	-1.117
$>\text{C}<$	-2.111
$=\text{CH}_2$	1.371
$=\text{CH}-$	0.284
$=\text{C}<$	-0.806
$=\text{C}=$	-0.966
$\equiv\text{CH}$	1.932
$\equiv\text{C}-$	0.773
<b>Halogen Increments</b>	
$-\text{F}$	-0.255
$-\text{Cl}$	0.902
$-\text{Br}$	1.094
$-\text{I}$	1.473
<b>Oxygen Increments</b>	
$-\text{OH}$ (alcohol)	-0.007
$-\text{O}-$ (nonring)	-0.854
$>\text{CO}$ (nonring)	-1.076
$-\text{CHO}$ (aldehyde)	0.297
$-\text{COOH}$ (acid)	-1.453
$=\text{O}$ (except for above)	-0.395
<b>Nitrogen Increments</b>	
$-\text{NH}_2$	1.383
$>\text{NH}$ (nonring)	0.487
$>\text{N}-$ (nonring)	-0.421
$-\text{CN}$	1.970
$-\text{NO}_2$	1.150
<b>Sulfur Increments</b>	
$-\text{SH}$	1.248
$-\text{S}-$ (nonring)	0.255

### A.9.3 $F_2$ Assumption

Joback's[61] factor analytic study found that the nine physical properties:

$$\begin{array}{lll}
 V_c & 1/\sqrt{P_c} & C_{p,298}^\circ \\
 n_A & \Delta G_{f,298}^\circ & \Delta H_{f,298}^\circ \\
 T_b & T_c & \Delta H_{vb}
 \end{array}$$

were related to three factors. These nine physical properties were weighted differently on each of the three factors.  $V_c$ ,  $1/\sqrt{P_c}$ ,  $C_{p,298}^\circ$ , and  $n_A$  were all heavily weighted on factor 1.  $\Delta G_{f,298}^\circ$  and  $\Delta H_{f,298}^\circ$  were heavily weighted on factor 2.  $T_b$ ,  $T_c$ , and  $\Delta H_{vb}$  were heavily weighted on factor 3. For many physical properties, e.g.,  $P_{vp}$ ,  $H_v$ ,  $\omega$ , the influence of factor 2 is least important. I assume an average value for factor 2 at times. The factor analysis technique used produced factors with zero means. The assumption I make is:

$$F_2 = 0.0 \quad (\text{A.27})$$

### A.9.4 $F_3$ Group Contribution Technique

Joback[61] proposed a group contribution technique for estimating  $F_3$ . I modified this technique to include some additional groups. The model for the technique is given by:

$$F_3 = -1.067 + \sum_{\text{all groups}} n_i \Delta_{i,F_3}. \quad (\text{A.28})$$

The values for the group contributions are given in Table A.8.

**Example:** Estimating  $F_3$  for 2,2-dimethylpentane:

Table A.8:  $F_3$  Group Contributions

Groups	Contribution
<b>Acyclic Increments</b>	
$-\text{CH}_3$	1.247
$>\text{CH}_2$	-0.226
$>\text{CH}-$	-1.481
$>\text{C}<$	-2.296
$=\text{CH}_2$	1.746
$=\text{CH}-$	-0.348
$=\text{C}<$	-1.848
$=\text{C}=$	-0.784
$\equiv\text{CH}$	1.506
$\equiv\text{C}-$	-0.489
<b>Halogen Increments</b>	
$-\text{F}$	1.807
$-\text{Cl}$	0.734
$-\text{Br}$	0.223
$-\text{I}$	-0.381
<b>Oxygen Increments</b>	
$-\text{OH}$ (alcohol)	-0.454
$-\text{O}-$ (nonring)	-0.074
$>\text{CO}$ (nonring)	-1.154
$-\text{CHO}$ (aldehyde)	0.398
$-\text{COOH}$ (acid)	0.196
$=\text{O}$ (except for above)	1.178
<b>Nitrogen Increments</b>	
$-\text{NH}_2$	1.094
$>\text{NH}$ (nonring)	-0.120
$>\text{N}-$ (nonring)	-1.224
$-\text{CN}$	-0.115
$-\text{NO}_2$	-0.057
<b>Sulfur Increments</b>	
$-\text{SH}$	0.269
$-\text{S}-$ (nonring)	-1.097

Group	Occurrences	Contribution	Total
-CH <sub>3</sub>	4	1.247	4.988
-CH <sub>2</sub> -	2	-0.226	-0.452
>C<	1	-2.296	-2.296
Total Contributions:			2.24

$$F_3 = -1.067 + 2.24 = 1.173. \quad (\text{A.29})$$

The literature value is reported as 0.823[61].

## A.10 Ideal Gas Heat Capacity

Typically ideal gas heat capacity is modeled as a polynomial in temperature. In my thesis I used a group contribution technique based on such a polynomial fit and an equation oriented technique derived from a factor analytic study of physical properties to estimate  $C_{pV}^\circ$  at 298K. The equation oriented technique did not account for variations in temperature.

### A.10.1 Group Contribution Technique

The ideal gas heat capacity is modeled as a cubic in temperature:

$$C_{pV}^\circ = C_{p,a} + C_{p,b}T + C_{p,c}T^2 + C_{p,d}T^3. \quad (\text{A.30})$$

Each of the coefficients of the polynomial is estimated by group contributions.  $C_{pV}^\circ$  has units of J/g-mol·K.  $C_{p,a}$ ,  $C_{p,b}$ ,  $C_{p,c}$ , and  $C_{p,d}$  have units J/g-mol·K, J/g-mol·K<sup>2</sup>,

J/g-mol·K<sup>3</sup>, and J/g-mol·K<sup>4</sup> respectively. Table A.9 presents the group contribution values.

**Example:** Estimating  $C_{p_v}^\circ$  for 2-nitrobutane at 298 and 800 K:

Group	Occurrences	Contributions			
		$C_{p,a}$	$C_{p,b}$	$C_{p,c}$	$C_{p,d}$
−CH <sub>3</sub>	2	1.95e+1	−8.08e−3	1.53e−4	−9.67e−8
−CH <sub>2</sub> −	1	−9.09e−1	9.50e−2	−5.44e−5	1.19e−8
>CH−	1	−2.30e+1	2.04e−1	−2.65e−4	1.20e−7
−NO <sub>2</sub>	1	2.59e+1	−3.74e−3	1.29e−4	−8.88e−8
Totals:		4.10e+1	2.79e−1	1.16e−4	−1.50e−7

$$C_{p_v}^\circ = 4.10e+1 + 2.79e-1 T + 1.16e-4 T^2 - 1.50e-7 T^3 \quad (A.31)$$

The estimated value for  $C_{p_v}^\circ$  at 298 K is 130.5 J/g-mol·K. The literature value is 123.55 J/g-mol·K[102]. The estimated value for  $C_{p_v}^\circ$  at 800 K is 261.6 J/g-mol·K. The literature value is 248.86 J/g-mol·K[102].

## A.10.2 Equation Oriented Technique

In a factor analytic study of physical properties  $C_{pV,298}^\circ$  was found to be related to three factors[61]:

$$C_{pV,298}^\circ = 25.70 - 8.72F_1 - 0.34F_2 - 2.44F_3. \quad (A.32)$$

Each of the three factors are estimated by group contribution techniques described in Section A.9.

Table A.9:  $C_{p_v}^o$  Group Contributions

Group		Contributions			
		$C_{p,a}$	$C_{p,b}$	$C_{p,c}$	$C_{p,d}$
<b>Acyclic Increments</b>					
-CH <sub>3</sub>		1.95e+1	-8.08e-3	1.53e-4	-9.67e-8
>CH <sub>2</sub>		-9.09e-1	9.50e-2	-5.44e-5	1.19e-8
>CH-		-2.30e+1	2.04e-1	-2.65e-4	1.20e-7
>C<		-6.62e+1	4.27e-1	-6.41e-4	3.01e-7
=CH <sub>2</sub>		2.36e+1	-3.81e-2	1.72e-4	-1.03e-7
=CH-		-8.00e+0	1.05e-1	-9.63e-5	3.56e-8
=C<		-2.81e+1	2.08e-1	-3.06e-4	1.46e-7
=C=		2.74e+1	-5.57e-2	1.01e-4	-5.02e-8
≡CH		2.45e+1	-2.71e-2	1.11e-4	-6.78e-8
≡C-		7.87e+0	2.01e-2	-8.33e-6	1.39e-9
<b>Ring Increments</b>					
>CH <sub>2</sub>		-6.03e+0	8.54e-2	-8.00e-6	-1.80e-8
>CH-		-2.05e+1	1.62e-1	-1.60e-4	6.24e-8
>C<		-9.09e+1	5.57e-1	-9.00e-4	4.69e-7
=CH-		-2.14e+0	5.74e-2	-1.64e-6	-1.59e-8
=C<		-8.25e+0	1.01e-1	-1.42e-4	6.78e-8
<b>Halogen Increments</b>					
-F		2.65e+1	-9.13e-2	1.91e-4	-1.03e-7
-Cl		3.33e+1	-9.63e-2	1.87e-4	-9.96e-8
-Br		2.86e+1	-6.49e-2	1.36e-4	-7.45e-8
-I		3.21e+1	-6.41e-2	1.26e-4	-6.87e-8
<b>Oxygen Increments</b>					
-OH (alcohol)		2.57e+1	-6.91e-2	1.77e-4	-9.88e-8
-O- (nonring)		2.55e+1	-6.32e-2	1.11e-4	-5.48e-8
-O- (ring)		1.22e+1	-1.26e-2	6.03e-5	-3.86e-8
>CO (nonring)		6.45e+0	6.70e-2	-3.57e-5	2.86e-9
>CO (ring)		3.04e+1	-8.29e-2	2.36e-4	-1.31e-7
-CHO (aldehyde)		3.09e+1	-3.36e-2	1.60e-4	-9.88e-8
-COOH (acid)		2.41e+1	4.27e-2	8.04e-5	-6.87e-8
-COO- (ester)		2.45e+1	4.02e-2	4.02e-5	-4.52e-8
=O (except for above)		6.82e+0	1.96e-2	1.27e-5	-1.78e-8

Table A.9 Continued:  $C_{pV}^o$  Group Contributions

Group	Contributions			
	$C_{p,a}$	$C_{p,b}$	$C_{p,c}$	$C_{p,d}$
Nitrogen Increments				
$-\text{NH}_2$	2.69e+1	-4.12e-2	1.64e-4	-9.76e-8
$>\text{NH}$ (nonring)	-1.21e+0	7.62e-2	-4.86e-5	1.05e-8
$>\text{NH}$ (ring)	1.18e+1	-2.30e-2	1.07e-4	-6.28e-8
$>\text{N-}$ (nonring)	-3.11e+1	2.27e-1	-3.20e-4	1.46e-7
$=\text{N-}$ (ring)	8.83e+0	-3.84e-3	4.35e-5	-2.60e-8
$=\text{NH}$	5.69e+0	-4.12e-3	1.28e-4	-8.88e-8
$-\text{CN}$	3.65e+1	-7.33e-2	1.84e-4	-1.03e-7
$-\text{NO}_2$	2.59e+1	-3.74e-3	1.29e-4	-8.88e-8
Sulfur Increments				
$-\text{SH}$	3.53e+1	-7.58e-2	1.85e-4	-1.03e-7
$-\text{S-}$ (nonring)	1.96e+1	-5.61e-3	4.02e-5	-2.76e-8
$-\text{S-}$ (ring)	1.67e+1	4.81e-3	2.77e-5	-2.11e-8

## A.11 Liquid Heat Capacity

There are three liquid heat capacities in common use,  $C_{pL}$ ,  $C_{\sigma L}$ , and  $C_{satL}$  [101].  $C_{pL}$  represents the change in enthalpy with temperature at constant pressure.  $C_{\sigma L}$  represents the change in enthalpy of a *saturated* liquid with temperature.  $C_{satL}$  represents the energy required to effect a temperature change while maintaining the liquid in a saturated state. The three heat capacities are related as follows:

$$C_{\sigma L} = C_{pL} + \left[ V_{\sigma L} - T \left( \frac{V}{T} \right)_p \right] \left( \frac{dP}{dT} \right)_{\sigma L} \quad (\text{A.33})$$

$$= C_{satL} + V_{\sigma L} \left( \frac{dP}{dT} \right)_{\sigma L} \quad (\text{A.34})$$

In my thesis I use an equation oriented technique to estimate  $C_{pL}$ .

### A.11.1 Equation Oriented Technique

Rowlinson[108] proposed an equation oriented estimation technique for  $C_{pL}$  which was later modified by Bondi[12]. The equation oriented technique models  $C_{pL}$  as a function of  $C_{pV}^o$ ,  $\omega$ , and  $T_c$ . The model is given by:

$$\frac{C_{pL} - C_p^o}{R} = 2.56 + 0.436(1 - T_r)^{-1} + \omega [2.91 + 4.28(1 - T_r)^{-1/3} T_r^{-1} + 0.296(1 - T_r)^{-1}] \quad (\text{A.35})$$

**Example:** Estimating the liquid heat capacity for acetone at 180, 209, and 297 K:

Values for the required properties are:  $T_c = 508.1$  K,  $\omega = 0.309$ . I estimated the values for  $C_{pV}^o$  at the three temperatures using the equation[101]:

$$C_{pV}^o = 1.505 + 6.224 \times 10^{-2} T - 2.992 \times 10^{-5} T^2 + 4.867 \times 10^{-9} T^3. \quad (\text{A.36})$$

The estimates and literature values are:

Temperature	$C_{pV}^o$	Estimate	Literature
180	11.77	26.68	28.0
209	13.25	27.26	28.2
297	17.48	30.23	29.8

Temperature is in K. Heat capacities are in cal/g-mol·K.

### A.12 Glass Transition Temperature

If a polymeric glass is heated it will begin to soften in the neighborhood of the glass transition temperature. On further heating the elastic behavior diminishes, but it is

only at temperatures more than 50° above  $T_g$  that a shear stress will cause viscous flow to predominate over elastic deformation[127]. I use a group contribution technique to estimate  $T_g$ .

### A.12.1 Group Contribution Technique

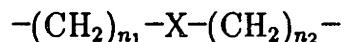
Van Krevelen and Hoftyzer[128] developed a group contribution technique to estimate  $T_g$ . The technique defined a *molar glass transition function*,  $Y_g$ :

$$Y_g = T_g M \quad (\text{A.37})$$

where  $M$  is the molecular weight of the polymeric repeat unit.  $Y_g$  is estimated from group contributions using the model:

$$Y_g = \sum_{\substack{\text{all} \\ \text{groups}}} Y_{gi}. \quad (\text{A.38})$$

Van Krevelen and Hoftyzer's estimation technique uses several *correction* terms for polar groups[127]. The interaction of polar groups is accounted for by means of an *interaction factor*,  $I_x$ .  $I_x$  represents the "linear concentration" of polar groups within a flexible chain of methylene beads.  $I_x$  is defined as the number of main chain atoms in the polar group,  $X$ , divided by the number of chain atoms of this group plus those of the directly connected methylene chains. For the configuration:



in which the characteristic group  $X$  contains  $n_x$  chain atoms, the formula of  $I_x$  is:

$$I_x = \frac{n_x}{n_x + n_1 + n_2}. \quad (\text{A.39})$$

$I_x$  is used to adjust the group contribution for the polar groups:



Their group contributions are:

$$Y_g(-\text{COO}-) = 8000 + 12000 I_x \quad (\text{A.40})$$

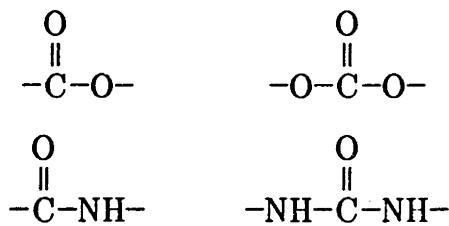
$$Y_g(-\text{OCOO}-) = 16000 + 10000 I_x \quad (\text{A.41})$$

$$Y_g(-\text{CONH}-) = 12000 + 1800 I_x^{-1} + 2 \times 10^6 \frac{n_\phi}{M} \quad (\text{A.42})$$

$$Y_g(-\text{NH}-\text{CO}-\text{NH}-) = 20000 + 2100 I_x^{-1} \quad (\text{A.43})$$

The manner in which  $I_x$  adjusts the contribution is thus dependent upon the particular polar group. My design procedures can not use these correction terms. I do not include them. Table A.10 shows the group contributions used in my thesis.

Van Krevelen[127] reported that in estimates of  $T_g$  for 600 polymers 80% of the errors were less than 20°K. This was for the unmodified group contribution technique. The modified estimation technique should have similar errors except for polymers containing:



For polymers containing these groups the errors will be greater.

**Example:** Estimating the glass transition temperature for poly(ethylene terephthalate)[127]:

Table A.10:  $Y_g$  Group Contributions

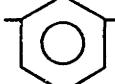
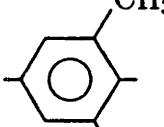
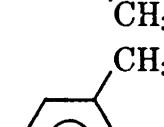
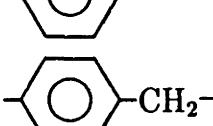
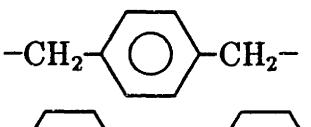
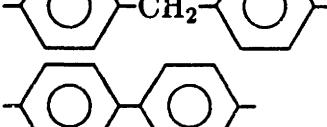
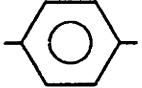
Groups	Contribution $\times 10^{-3}$
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	2.7
$-\text{CH}(\text{CH}_3)-$	8.0
$-\text{CH}(\text{C}_6\text{H}_5)-$	35.0
$-\text{C}(\text{CH}_3)_2-$	15.0
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	50.0
	32.0
	28.0
	7.0
	55.0
	35.0
	29.8
	37.4
	79.7
	77.0

Table A.10 Continued:  $Y_g$  Group Contributions

Groups	Contribution $\times 10^{-3}$
<b>Bifunctional Oxygen Groups</b>	
-O-	4.0
-CO-	27.0
-COO-	8.0
-O-COO-	16.0
-CO-O-CO-	20.0
-CH(OH)-	13.0
-O-CH <sub>2</sub> -O-	10.7
<b>Bifunctional Nitrogen and Oxygen Groups</b>	
-CO-NH-	12.0
-O-CO-NH-	25.0
-NH-CO-NH-	20.0
<b>Bifunctional Halogen Groups</b>	
-CHF-	11.0
-CF <sub>2</sub> -	13.0
-CHCl-	20.0
-CCl <sub>2</sub> -	25.0
-CFCl-	23.0

Group	Occurrences	Contribution	Total
$-\text{CH}_2-$	2	2700	5400
	1	32000	32000
$-\text{COO}-$	2	8000	16000
Total Contributions: 53400.			

The molecular weight of the polymeric repeat unit is 192.17 yielding a  $T_g$  of 278K. The literature value ranges from 342 to 350K[127].

## A.13 Gas Permeability

The permeability relationship for non-swelling gases and non-interacting liquids is:

$$P = Dk \quad (\text{A.44})$$

where  $P$  = permeability,  $D$  = diffusion, and  $k$  = solubility of permeant in polymer. Two techniques are available for the estimation of gas permeability through a polymer[110, 74]. I used Salame's group contribution technique[110] in my thesis.

### A.13.1 Group Contribution Technique

Salame[110] defines a parameter  $\pi$ , called permachor, which is related to the permeability of a polymer by:

$$P = Ae^{-S\pi} \quad (\text{A.45})$$

$A$  and  $S$  are constants for the gas permeant. The values for  $A$  and  $S$  for the three permeants oxygen, nitrogen, and carbon dioxide are shown in Table A.11.

Table A.11: *A* and *S* Parameters for Permачор Estimation

Permeant	<i>A</i> <sup>†</sup>	<i>S</i>
Oxygen	8850	0.112
Nitrogen	3000	0.121
Carbon Dioxide	55100	0.122

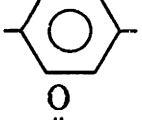
† The units for *A* are:  $\left( \frac{\text{cc-mil}}{100 \text{ in}^2 \cdot \text{day-Atm}} \right)$ . *S* is dimensionless.

$\pi$  is estimated by a group contribution technique. The model for the technique is:

$$\pi = \left( \sum_{\text{all groups}} n_i \Delta_{i,\pi} \right) / N \quad (\text{A.46})$$

where *N* is the number of groups in the backbone repeat unit. The group contributions for  $\pi$  are given in Table A.12.

**Example:** Estimating the permeability of oxygen through polycarbonate at 25°C.

Group	Occurrences	Contribution	Total
$-\text{CH}_3$	2	15	30
$>\text{C}<$	1	-50	-50
	2	60	120
$-\text{C}=\text{O}-$	1	24	24

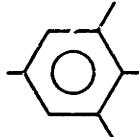
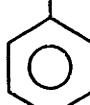
Total Contributions: 124.

$$\pi = 124/4 = 31 \quad (\text{A.47})$$

The estimated permeability is thus:

$$P = 8850 e^{(-0.112)(31)} = 275 \text{ cc-mil}/100 \text{ in}^2 \cdot \text{day-Atm.} \quad (\text{A.48})$$

Table A.12:  $\pi$  Group Contributions

Groups	Contribution
<b>Backbone Segments</b>	
$-\text{CH}_2-$	15
$>\text{CH}-$	0
$>\text{C}<$	-50
$-\text{CH}=\text{CH}-$	-12
	60
	-74
$-\text{O}-$	70
$-\text{COO}-$	102
$-\text{O}-\text{COO}-$	24
$-\text{CO}-\text{NH}-$	309 <sup>†</sup>
<b>Side Chain Substituents</b>	
$-\text{CH}_3$	15
$-\text{CH}_2\text{CH}(\text{CH}_3)-$	-1
	39
$-\text{Cl}$	108
$-\text{OH}$	255 <sup>†</sup>
$-\text{CH}_2\text{Cl}$	50
$-\text{CN}$	205
$-\text{F}$	85

<sup>†</sup> Dry value used.

Table A.13: Gas Permeability Estimates

Polymer	$P_{O_2}$ at 25°C (cc-mil/100in <sup>2</sup> -day-atm)	
	Estimate	Literature
Polybutadiene	4.52e+3	3.2e+3
Polyethylene	1.65e+3	3.0e+3 <sup>†</sup>
Polycarbonate	2.75e+2	3.0e+2
Polystyrene	4.30e+2	3.5e+2
Polyethylene terephthalate	1.19e+1	2.4e+1 <sup>†</sup>
Polymethyl methacrylate	2.36e+2	1.7e+1
Polyvinyl chloride	9.50e+0	7.0e+0
Polyacrylonitrile	3.90e-2	2.5e-2

<sup>†</sup> Original data was on partially crystalline polymers. The data was adjusted to account for crystallinity.

The literature value is 300 cc-mil/100 in<sup>2</sup>-day-Atm[74]. Table A.13 shows comparisons between estimates and literature values of the permeability of oxygen for additional polymers.

## A.14 Volume Resistivity

The electrical resistance of most polymers is very high and conductivity probably results from the presence of ionic impurity, whose mobility is limited by the very high viscosity of the medium. The conductivity and the activation energy of conduction appear to be practically insensitive to crystallinity[127]. I use an equation oriented technique to estimate the volume resistance,  $R$ .

### A.14.1 Equation Oriented Technique

Van Krevelen[127] identified a linear relationship between the log of  $R$  and the dielectric constant,  $\epsilon$ . The model of this relationship is given by:

$$\log R = 23 - 2\epsilon. \quad (\text{A.49})$$

The units of  $R$  are  $\Omega\cdot\text{cm}$ . The model was developed from data taken at 298K.

## A.15 Molar Polarization

Applying an electric field to a material may cause charge to flow within the material or may produce finite changes in the relative positions of the electric charges in the material. If charge flows upon application of an electric field the material is a *conductor*. If electric charges are displaced the material is a *dielectric*. All common polymers are dielectrics[127].

The dielectric constant,  $\epsilon$ , of insulating materials is the ratio of the capacities of a parallel plate condenser measured with and without the dielectric material placed between the plates. The difference is due to the polarization of the dielectric. The molar polarization of a dielectric is defined as:

$$P_{LL} = \frac{\epsilon - 1}{\epsilon + 2} V \quad (\text{A.50})$$

where  $V$  is the molar volume. I estimate  $P_{LL}$  using a group contribution technique.

### A.15.1 Group Contribution Technique

Van Krevelen[127] developed a group contribution technique to estimate  $P_{LL}$ . The model for the technique is given by:

$$P_{LL} = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,P_{LL}}. \quad (\text{A.51})$$

The values for the group contributions are given in Table A.14.

**Example:** Estimating the molar polarization of polycarbonate:

Group	Occurrences	Contribution	Total
$-\text{CH}_3$	2	5.64	11.28
$>\text{C}<$	1	2.58	2.58
	2	25.0	50.0
$-\text{O}-\text{C}=\text{O}-$	1	22	22
Total Contributions: 85.86			

The estimated value of  $P_{LL}$  is 85.86  $\text{cm}^3/\text{g}\cdot\text{mol}$ . The molar volume at 298K is 215  $\text{cm}^3/\text{mol}$ . We estimate the dielectric constant to be 3.0. This is in adequate agreement with the literature value of 2.6–3.0[127].

### A.16 Molar Volume

Molar volume is regarded as one of the most important polymer physical properties[127].

Molar volume is useful in characterizing polymers and is used in numerous physical

Table A.14:  $P_{LL}$  Group Contributions

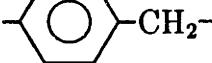
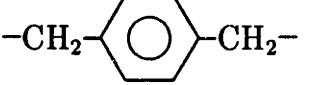
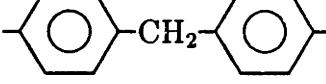
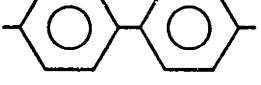
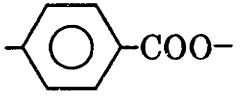
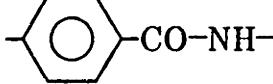
Groups	Contribution
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	4.65
$-\text{CH}(\text{CH}_3)-$	9.26
$-\text{CH}(\text{C}_6\text{H}_5)-$	29.12
$-\text{C}(\text{CH}_3)_2-$	13.86
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	33.72
	25.0
	29.65
	34.3
	54.65
	50.0
Bifunctional Oxygen Groups	
$-\text{O}-$	5.2
$-\text{CO}-$	10.0
$-\text{COO}-$	15.0
$-\text{O}-\text{COO}-$	22.0
$-\text{CO}-\text{O}-\text{CO}-$	25.0
$-\text{CH}(\text{OH})-$	10.0

Table A.14 Continued:  $P_{LL}$  Group Contributions

Groups	Contribution
Bifunctional Oxygen Groups continued	
	40.0
$-\text{O}-\text{CH}_2-\text{O}-$	15.05
Bifunctional Nitrogen Groups	
$-\text{NH}-$	14.6
Bifunctional Nitrogen and Oxygen Groups	
$-\text{CO}-\text{NH}-$	30.0
	55.0
Bifunctional Halogen Groups	
$-\text{CHF}-$	5.42
$-\text{CF}_2-$	6.25
$-\text{CHCl}-$	13.7
$-\text{CCl}_2-$	17.7
$-\text{CFCl}-$	13.9

property correlations. I use a group contribution technique to estimate the molar volume.

### A.16.1 Group Contribution Technique

Group contribution estimation techniques are available for both the molar volume of rubbery amorphous polymers and glassy amorphous polymers[127]. The models for both these techniques are:

$$V_g = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,V_g} \quad (\text{A.52})$$

$$V_r = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,V_r} \quad (\text{A.53})$$

These contributions are presented in Table A.15.

Table A.16 compares estimated molar volumes for rubbery polymers against literature values[127]. In general the estimates are quite good. The average deviation from literature values for 40 estimates was 1.5%[127].

Table A.17 compares estimated molar volumes for glassy polymers against literature values[127]. In general the estimates are quite good. The average percent error for 67 estimates was 1.2%[127].

### Averaged Contributions

Examining the contributions for any group toward  $V_r$  and  $V_g$  in Table A.15 we see that the difference between the contributions is small. Since in the design of molecules we do not know ahead of time whether the final molecule designed will be rubbery or glassy I averaged these group contributions to yield a set of molar volume contributions. These

Table A.15: Molar Volume Group Contributions

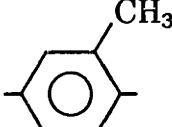
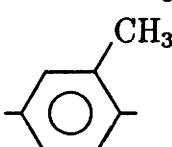
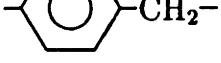
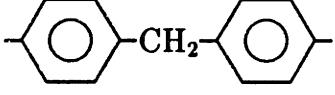
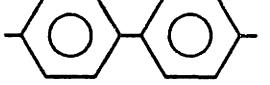
Groups	$V_g$ Contribution	$V_r$ Contribution
Bifunctional Hydrocarbon Groups		
$-\text{CH}_2-$	15.85	16.45
$-\text{CH}(\text{CH}_3)-$	33.35	32.65
$-\text{CH}(\text{C}_6\text{H}_{11})-$	100.15	—
$-\text{CH}(\text{C}_6\text{H}_5)-$	82.15	74.5
$-\text{C}(\text{CH}_3)_2-$	52.4	50.35
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	100.20	92.2
$-\text{CH}=\text{CH}-$	—	27.75
$-\text{CH}=\text{C}(\text{CH}_3)-$	—	42.8
	65.5	61.4
	104.1	—
	83.4	—
	81.35	77.85
$-\text{CH}_2-\text{C}_6\text{H}_4-\text{CH}_2-$	97.20	94.30
	146.85	139.25
	131.0	122.8

Table A.15 Continued: Molar Volume Group Contributions

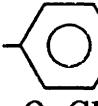
Groups	$V_g$ Contribution	$V_r$ Contribution
<b>Bifunctional Oxygen Groups</b>		
-O-	10.0	8.5
-CO-	13.4	-
-COO-	20.63	22.8
-O-COO-	31.4	-
-CH(OH)-	19.15	-
 -COO-	88.5	86.0
-O-CH <sub>2</sub> -O-	35.85	33.45
<b>Bifunctional Nitrogen Groups</b>		
-CH(CN)-	28.95	-
<b>Bifunctional Nitrogen and Oxygen Groups</b>		
-CO-NH-	24.9	-
 -CO-NH-	90.4	-
<b>Bifunctional Sulfur Groups</b>		
-S-	17.8	15.0
-S-S-	35.6	30.0
-S-CH <sub>2</sub> -S-	51.45	46.45
<b>Bifunctional Halogen Groups</b>		
-CHF-	20.35	19.85
-CF <sub>2</sub> -	26.4	24.75
-CHCl-	29.35	28.25
-CCl <sub>2</sub> -	44.4	41.55
-CH=CCl-	-	38.4
-CFCl-	35.4	33.15
-CHBr-	39.0	-
-CBr <sub>2</sub> -	46.0	-

Table A.16: Molar Volumes of Rubbery Amorphous Polymers at 25°C

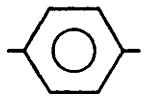
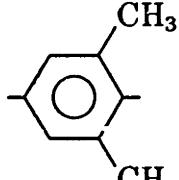
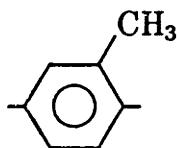
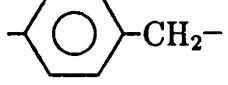
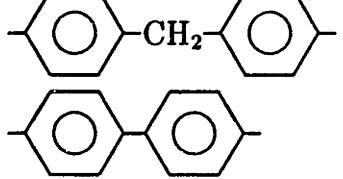
Polymer	$V_r$ cm <sup>3</sup> /mol	
	Literature	Experimental
polyethylene	32.8	32.9
polypropylene	49.5	49.1
polybutene	65.2	65.8
poly(tetrafluoroethylene)	50.0	49.5
poly(methyl methacrylate)	70.6	70.1
polybutadiene	60.7	60.7
polyacetaldehyde	41.2	41.2
poly(propylene oxide)	58.1	57.6

Table A.17: Molar Volumes of Glassy Amorphous Polymers at 25°C

Polymer	$V_g$ cm <sup>3</sup> /mol	
	Literature	Experimental
polystyrene	100.2	98.4
poly(vinyl chloride)	45.1	45.2
poly(vinyl alcohol)	35.0	35.0
poly(vinyl acetate)	72.4	72.2
poly(methyl methacrylate)	85.6	86.5
polyacrylonitrile	44.8	44.8
poly(ethylene terephthalate)	144.5	143.2
nylon 6	104.4	104.2

contributions are show in Table A.18. Some groups had contributions for only  $V_g$  or  $V_r$ . In these cases I chose the available contribution to represent the molar volume contribution.

Table A.18: Molar Volume Group Contributions

Groups	Contribution
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	16.15
$-\text{CH}(\text{CH}_3)-$	33.00
$-\text{CH}(\text{C}_6\text{H}_{11})-$	100.15
$-\text{CH}(\text{C}_6\text{H}_5)-$	78.32
$-\text{C}(\text{CH}_3)_2-$	51.38
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	96.20
$-\text{CH}=\text{CH}-$	27.75
$-\text{CH}=\text{C}(\text{CH}_3)-$	42.80
	63.45
	104.10
	83.40
	79.60
	95.75
	143.05
	126.90

**Example:** Estimating the molar volume of poly(ethylene terephthalate):

Group	Occurrences	Contribution	Total
$-\text{CH}_2-$	2	16.15	32.3
	1	63.45	63.45
$-\text{O}-\text{C}=$	2	21.72	43.44
Total Contributions: 139.19			

The estimated value of  $V$  is  $139.19 \text{ cm}^3/\text{g-mol}$ . The literature value for  $V_g$  is  $144.5 \text{ cm}^3/\text{g-mol}$ .

## A.17 Molecular Weight

Molecular weight is the single physical property for which group contributions give an exact value. Table A.19 lists the molecular weight contributions for a number of polymer groups[127].

## A.18 Polymer Thermal Conductivity

No adequate theory exists which may be used to predict the thermal conductivity of polymeric melts or solids[127]. Most of the theoretical or semi-theoretical expressions proposed are based on Debye's[24] treatment of heat conductivity which lead to:

$$\lambda = \Lambda c_V \rho u L \quad (\text{A.54})$$

Table A.18 Continued: Molar Volume Group Contributions

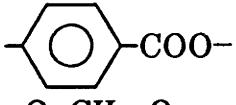
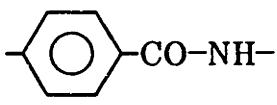
Groups	Contribution
<b>Bifunctional Oxygen Groups</b>	
-O-	9.25
-CO-	13.40
-COO-	21.72
-O-COO-	31.40
-CH(OH)-	19.15
	87.25
-O-CH <sub>2</sub> -O-	34.65
<b>Bifunctional Nitrogen Groups</b>	
-CH(CN)-	28.95
<b>Bifunctional Nitrogen and Oxygen Groups</b>	
-CO-NH-	24.9
	90.4
<b>Bifunctional Halogen Groups</b>	
-CHF-	16.4
-CF <sub>2</sub> -	25.58
-CHCl-	28.8
-CCl <sub>2</sub> -	42.98
-CH=CCl-	38.4
-CFCl-	34.28
-CHBr-	39.0
-CBr <sub>2</sub> -	46.0

Table A.19:  $M_w$  Group Contributions

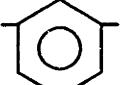
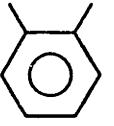
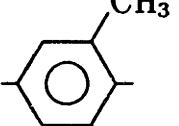
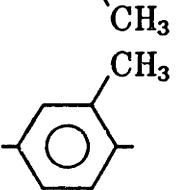
Groups	Contribution
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	14.03
$-\text{CH}(\text{CH}_3)-$	28.05
$-\text{CH}(\text{C}_6\text{H}_5)-$	90.12
$-\text{C}(\text{CH}_3)_2-$	42.08
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	104.14
	76.09
	76.09
	76.09
	104.14
	90.12
Bifunctional Oxygen Groups	
$-\text{O}-$	16.00
$-\text{CO}-$	28.01
$-\text{CH}(\text{OH})-$	30.03

Table A.19 Continued:  $M_w$  Group Contributions

Groups	Contribution
Bifunctional Nitrogen Groups	
-NH-	15.02
Bifunctional Halogen Groups	
-CHF-	32.02
-CF <sub>2</sub> -	50.01
-CHCl-	48.48
-CCl <sub>2</sub> -	82.92
-CFCl-	66.47

where  $c_V$  is the specific heat capacity,  $\rho$  is density,  $u$  is the velocity of sound,  $L$  is average free path length, and  $\Lambda$  is a constant. I use an equation oriented technique based on Equation A.54 to estimate thermal conductivity.

### A.18.1 Equation Oriented Technique

Van Krevelen[127] developed an equation oriented technique to estimate the thermal conductivity of amorphous polymers and polymer melts.  $\lambda$  is modeled as a function of the solid heat capacity, the molar volume, and the Rao function which is a function of the velocity of sound. The model is given by:

$$\lambda(298) = L \left( \frac{C_p}{V} \right) \left( \frac{U}{V} \right)^3 \quad (A.55)$$

where  $L \approx 5 \times 10^{-11}$  m.  $\lambda$  has units of (J/s·m·K).

The three required properties,  $C_p$ ,  $V$ , and  $U$ , are all fundamental properties. The estimation of  $C_p$  is described in Section A.19,  $V$  in Section A.16, and  $U$  in Section A.20.

Table A.20: Thermal Conductivities of Amorphous Polymers

Polymer	$\lambda$ (J/s·m·K)	
	Estimated	Literature
Polypropylene (at.)	0.147	0.172
Polyisobutylene	0.161	0.130
Polystyrene	0.140	0.142
Poly(vinyl chloride)	0.112	0.168
Poly(methyl methacrylate)	0.156	0.193
Poly(ethylene oxide)	0.193	0.205
Poly(ethylene terephthalate)	0.152	0.218

**Example:** Estimating the thermal conductivity of amorphous poly(methyl methacrylate) at 298K[127]:

Values for the required properties are:  $C_{ps} = 1.381 \times 10^5$  J/kg·mol·K,  $V = 0.0856$  m<sup>3</sup>/kg·mol,  $U = 1.073$  m<sup>10/3</sup>/s<sup>1/3</sup>·kg·mol.  $U$  was estimated by the group contribution technique discussed in Section A.20.

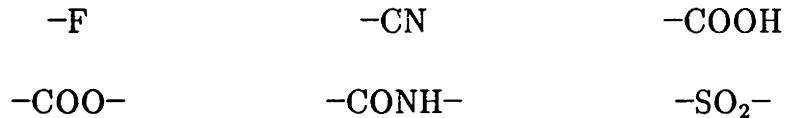
The estimated value for  $\lambda$  is 0.159 J/s·m·K. The literature value is 0.193 J/s·m·K[127]. Table A.20 presents a comparison of estimates and literature values for 7 amorphous polymers[127].

## A.19 Solid Molar Heat Capacity

The specific heat capacity is equal to the quantity of heat needed to raise one kilogram of a compound 1K. Data for the molar heat capacity of polymers in the solid state is limited. I use a group contribution technique to estimate  $C_{ps}$  in my thesis.

### A.19.1 Group Contribution Technique

Satoh[111] developed a group contribution technique to estimate  $C_{ps}$ . Van Krevelen[127] added group contributions for:



The model for the technique is given by:

$$C_{ps} = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,C_{ps}}. \quad (\text{A.56})$$

The group contributions are for  $C_{ps}$  are given in Table A.21.

**Example:** Estimating the solid heat capacity of poly(vinyl alcohol):

Group	Occurrences	Contribution	Total
$>CH-$	1	15.6	15.6
$-CH_2-$	1	25.35	25.35
$-OH$	1	17.0	17.0
Total Contributions:			57.95

The estimated value of  $C_{ps}$  is 57.95 J/kg-mol·K. The literature value is reported as 57J/kg-mol·K[127].

Table A.21:  $C_{ps}$  Group Contributions

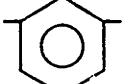
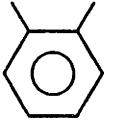
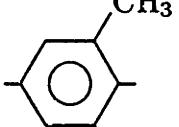
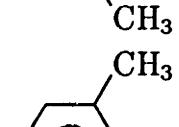
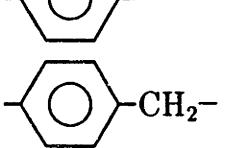
Groups	Contribution
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	25.35
$-\text{CH}(\text{CH}_3)-$	46.5
$-\text{CH}(\text{C}_5\text{H}_9)-$	110.8
$-\text{CH}(\text{C}_6\text{H}_11)-$	121.2
$-\text{CH}(\text{C}_6\text{H}_5)-$	101.2
$-\text{C}(\text{CH}_3)_2-$	68.0
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	122.7
$-\text{CH}=\text{CH}-$	37.3
$-\text{CH}=\text{C}(\text{CH}_3)-$	60.05
	78.8
	78.8
	78.8
	126.8
	102.75
	104.15

Table A.21 Continued:  $C_{ps}$  Group Contributions

Groups	Contribution
Bifunctional Hydrocarbon Groups continued	
	129.5
	182.95
	157.6
Bifunctional Oxygen Groups	
$-O-$	16.8
$-CO-$	23.05
$-COO-$	46.0
$-CH(OH)-$	32.6
$-CH(COOH)-$	65.6
	124.8
$-O-CH_2-O-$	58.95
Bifunctional Nitrogen-Oxygen Groups	
$-CO-NH-$	46.0
$-CH(NO_2)-$	57.5
	124.8
Bifunctional Halogen Groups	
$-CHF-$	37.0
$-CF_2-$	49.0
$-CHCl-$	42.7
$-CCl_2-$	60.4
$-CFCl-$	54.7

## A.20 Rao Function

Rao[97] demonstrated that for organic liquids the ratio  $u^{1/3}/\rho$  is practically independent of temperature and that the combination:

$$U = V u^{1/3} \quad (\text{A.57})$$

is an additive molar quantity.  $u$  is the velocity of sound through the material.  $U$  is called the *Rao function* or *molar sound velocity function*[127]. Schuyer[112] showed that the generalized Rao function is:

$$U = V u^{1/3} \left[ \frac{1 + \nu}{3(1 - \nu)} \right]. \quad (\text{A.58})$$

The Rao function is used in equation oriented techniques for estimating Poisson's ratio and the thermal conductivity. I use a group contribution technique to estimate  $U$ .

### A.20.1 Group Contribution Technique

Van Krevelen[127] presented a group contribution technique to estimate  $U$ . The model for the technique is:

$$U = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,U}. \quad (\text{A.59})$$

The group contributions are given in Table A.22.

**Example:** Estimating the Rao function for polycarbonate:

Table A.22: *U* Group Contributions

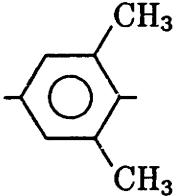
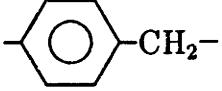
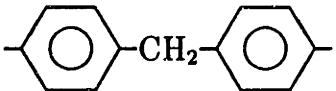
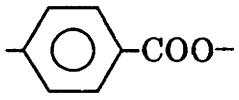
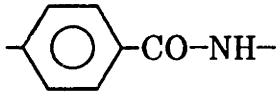
Groups	Contribution $\times 10^{-3}$
Bifunctional Hydrocarbon Groups	
$-\text{CH}_2-$	0.880
$-\text{CH}(\text{CH}_3)-$	1.850
$-\text{CH}(\text{C}_6\text{H}_5)-$	5.100
$-\text{C}(\text{CH}_3)_2-$	2.850
$-\text{C}(\text{CH}_3)(\text{C}_6\text{H}_5)-$	6.100
	4.100
	6.150
	4.980
$-\text{CH}_2-\text{C}_6\text{H}_5-\text{CH}_2-$	5.860
	9.080
	8.200

Table A.22 Continued: *U* Group Contributions

Groups	Contribution $\times 10^{-3}$
Bifunctional Oxygen Groups	
-O-	0.400
-CO-	0.900
-COO-	1.250
-O-COO-	1.600
-CO-O-CO-	2.150
-CH(OH)-	1.050
	5.350
-O-CH <sub>2</sub> -O-	1.680
Bifunctional Nitrogen and Oxygen Groups	
-CO-NH-	1.700
-O-CO-NH-	2.100
	5.800
Bifunctional Halogen Groups	
-CHF-	0.950
-CF <sub>2</sub> -	1.050
-CHCl-	1.600
-CCl <sub>2</sub> -	2.350
-CH=CCl-	1.900
-CFCl-	1.700

Group	Occurrences	Contribution	Total
$-\text{CH}_3$	2	1400	2800
$>\text{C}<$	1	50	50
	2	4100	8200
$-\text{O}-\text{C}=\text{O}-$	1	1600	1600
Total Contributions: 12650			

The estimated value of  $U$  is  $12650 \text{ cm}^{10/3}/\text{s}^{1/3} \cdot \text{g-mol}$ .

## A.21 Solubility Parameters

Hildebrand and Scott[57] assumed regular solution theory to relate a solute's activity to the pure component solubility parameters of the solution's components. Regular solution theory is not widely applicable being accurate for certain systems involving hydrocarbons and fluorocarbons[77]. Applicability is dependent upon the absence of permanent dipole-dipole interactions and hydrogen bonding. The solubility parameter was divided into three parameters to account for such interactions. The three parameters are:

$\delta_d$ : the dispersive forces solubility parameter.

$\delta_p$ : the polar forces solubility parameter.

$\delta_h$ : the hydrogen-bonding solubility parameter.

In my solvent design case study it was necessary to estimate  $\delta_p$  and  $\delta_h$ . Group contribution estimation techniques are available for these physical properties[115]. Unfortunately these techniques are nonlinear and are not capable of being linearized. I thus developed my own group contribution estimation techniques. A major effort was not given to make the estimation techniques highly accurate. The estimation techniques developed are considered sufficiently accurate to present the concepts of the case study.

### **A.21.1 $\delta_p$ Group Contribution Technique**

Polar solubility parameter values for 77 compounds were used in the development of the group contribution estimation technique. The data was taken from Barton[6]. Adequate data was available to determine contributions for 15 acyclic groups. The contributions for these groups are shown in Figure A.23.

The intercept of the multiple regression is  $-26.7$ . The model for the group contribution technique is thus:

$$\delta_p = \sum_{\text{all groups}} n_i \Delta_{i,\delta_p} - 26.7 \quad (\text{A.60})$$

The average absolute error of the regression is 0.57. The maximum absolute error was 2.37 associated with the estimation of 1-decene. Table A.24 shows the estimates and errors for 25 compounds not included in the regression. The average absolute error is 1.61. The maximum absolute error is 10.24 associated with the estimation of glycerol. Not accounting for the estimation of glycerol the average absolute error of the remaining 24 compounds is 1.33.

Table A.23:  $\delta_p$  Group Contributions

Groups	Contribution
$-\text{CH}_3$	1.47e+1
$-\text{CH}_2-$	-3.28e - 1
$>\text{CH}-$	-1.60e+1
$>\text{C}<$	-3.16e+1
$=\text{CH}_2$	1.61e+1
$=\text{CH}-$	5.73e - 1
$=\text{C}<$	-1.42e+1
$-\text{Cl}$	1.92e+1
$-\text{OH}$ (alcohol)	2.30e+1
$-\text{O-}$ (nonring)	1.11e+0
$>\text{CO}$ (nonring)	4.80e+0
$-\text{CHO}$ (aldehyde)	2.21e+1
$-\text{COOH}$ (acid)	2.39e+1
$-\text{COO-}$ (ester)	6.17e+0
$-\text{CN}$	2.32e+1

### A.21.2 $\delta_h$ Group Contribution Technique

Hydrogen solubility parameter values for 77 compounds were used in the development of the group contribution estimation technique. The data was taken from Barton[6]. Adequate data was available to determine contributions for 15 acyclic groups. The contributions for these groups are shown in Figure A.25.

The intercept of the multiple regression is 29.1. The model for the group contribution technique is thus:

$$\delta_h = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,\delta_h} + 29.1 \quad (\text{A.61})$$

The average absolute error of the regression is 1.08. The maximum absolute error was 5.42 associated with the estimation of methanol. Table A.26 shows the estimates and errors for 25 compounds not included in the regression. The average absolute

Table A.24: Example  $\delta_p$  Estimation Errors

Solvent	$\delta_p$ Literature	$\delta_p$ Estimate	Error <sup>†</sup>
Butyl Ether	4.30	1.84	-2.46
2,3-Dichloropropanol	13.60	17.40	3.80
3,3-Diethylpentane	0.00	-0.81	-0.81
2-Ethylhexyl Chloride	5.40	4.26	-1.14
Ethyl Propionate	8.10	8.21	0.11
Ethyl Isopropyl Ether	5.10	2.18	-2.92
Ethyl Butyrate	7.60	7.89	0.29
Ethylene Glycol	15.10	18.64	3.54
2-Chloroethanol	13.30	14.20	0.90
3-Ethylheptane	0.00	-0.24	-0.24
Isobutyl Acetate	7.80	7.24	-0.56
Isobutyraldehyde	9.00	8.80	-0.20
Isobutyric Acid	9.90	10.60	0.70
Isobutyl Heptyl Ketone	6.00	3.90	-2.10
Isopropyl Acetate	8.40	7.57	-0.83
Isobutyronitrile	10.60	9.90	-0.70
3-Methyl-1-Pentene	3.70	3.05	-0.66
2-Methylpropene	4.20	4.60	0.40
Methyl Propyl Ketone	8.80	6.84	-1.96
2-Octene,cis	3.20	2.53	-0.67
3-Ethyl-2,2-Dimethylpentane	0.00	-1.46	-1.46
Valeric Acid	9.70	10.92	1.22
Hexanoic Acid	9.20	10.59	1.39
Hexaldehyde	7.90	8.79	0.89
Glycerol	15.40	25.64	10.24

<sup>†</sup> Error = Estimate - Literature.

Table A.25:  $\delta_h$  Group Contributions

Groups	Contribution
$-\text{CH}_3$	-12.60
$-\text{CH}_2-$	-0.512
$>\text{CH}-$	-10.90
$>\text{C}<$	22.50
$=\text{CH}_2$	-12.60
$=\text{CH}-$	0.0974
$=\text{C}<$	12.80
$-\text{Cl}$	-10.30
$-\text{OH}$ (alcohol)	2.08
$-\text{O-}$ (nonring)	1.82
$>\text{CO}$ (nonring)	5.61
$-\text{CHO}$ (aldehyde)	-4.81
$-\text{COOH}$ (acid)	1.83
$-\text{COO-}$ (ester)	4.36
$-\text{CN}$	-1.22

error is 2.16. The maximum absolute error is 13.82 associated with the estimation of glycerol. Not accounting for the estimation of glycerol the average absolute error of the remaining 24 compounds is 1.30.

## A.22 Drug Design

Hansch and Leo[53] tabulated substituent constants for a large number of groups. For interactive design I needed group contributions for  $\pi$  and  $\sigma_M$ . The constants provided were not for an additive group contribution model. I took the values for their substituents and regressed them onto a set of groups assuming a linear group contribution model.

Table A.26: Example  $\delta_h$  Estimation Errors

Solvent	$\delta_h$ Literature	$\delta_h$ Estimate	Error <sup>†</sup>
Butyl Ether	4.50	2.65	-1.85
2,3-Dichloropropanol	16.40	19.78	3.38
3,3-Diethylpentane	0.00	-0.85	-0.85
2-Ethylhexyl Chloride	2.40	1.94	-0.46
Ethyl Propionate	7.80	7.24	-0.56
Ethyl Isopropyl Ether	6.20	3.51	-2.69
Ethyl Butyrate	6.40	6.72	0.32
Ethylene Glycol	29.80	32.24	2.44
2-Chloroethanol	19.60	19.18	-0.42
3-Ethylheptane	0.00	-0.36	-0.36
Isobutyl Acetate	5.10	6.05	0.95
Isobutyraldehyde	8.40	9.99	1.59
Isobutyric Acid	16.20	16.63	0.43
Isobutyl Heptyl Ketone	3.60	4.23	0.63
Isopropyl Acetate	5.70	6.56	0.86
Isobutyronitrile	10.60	13.58	2.98
3-Methyl-1-Pentene	2.90	1.79	-1.11
2-Methylpropene	3.00	4.10	1.10
Methyl Propyl Ketone	7.00	8.49	1.49
2-Octene,cis	0.00	2.05	2.05
3-Ethyl-2,2-Dimethylpentane	0.00	-1.52	-1.52
Valeric Acid	14.30	16.79	2.49
Hexanoic Acid	13.80	16.28	2.48
Hexaldehyde	16.90	9.64	-7.26
Glycerol	31.40	45.22	13.82

<sup>†</sup> Error = Estimate - Literature.

### A.22.1 $\sigma_M$

Constants for 87 substituents were regressed into contributions for 29 groups. The intercept of the regression was set to zero. The model of the group contribution technique is thus:

$$\sigma_M = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,\sigma_M} \quad (\text{A.62})$$

The regression had an  $R^2$  value of 0.81. The average absolute error of the regression was 0.063. The substituents had an average  $\sigma_M$  value of 0.236. Table A.27 shows the contributions.

### A.22.2 $\pi$

Constants for 68 substituents were regressed into contributions for 24 groups. The intercept of the regression was set to zero. The model of the group contribution technique is thus:

$$\pi = \sum_{\substack{\text{all} \\ \text{groups}}} n_i \Delta_{i,\pi} \quad (\text{A.63})$$

The regression had an  $R^2$  value of 0.90. The average absolute error of the regression was 0.225. The substituents had an average  $\pi$  value of 0.738. Table A.28 shows the contributions.

Table A.27:  $\sigma_M$  Group Contributions

Groups	Contribution
$-\text{CH}_3$	8.85e-2
$-\text{CH}_2-$	-6.10e-2
$>\text{CH}-$	-2.26e-1
$>\text{C}<$	-4.14e-1
$=\text{CH}_2$	1.19e-1
$=\text{CH}-$	-6.91e-2
$=\text{C}<$	-4.39e-1
$=\text{C}=$	-5.80e-1
$\equiv\text{CH}$	1.53e-1
$\equiv\text{C}-$	5.73e-2
$-\text{F}$	2.36e-1
$-\text{Cl}$	2.33e-1
$-\text{Br}$	2.41e-1
$-\text{I}$	2.56e-1
$-\text{OH}$ (alcohol)	1.06e-1
$-\text{O}-$ (nonring)	1.15e-1
$>\text{CO}$ (nonring)	2.78e-1
$-\text{CHO}$ (aldehyde)	3.63e-1
$-\text{COOH}$ (acid)	2.47e-1
$-\text{COO}-$ (ester)	2.88e-1
$=\text{O}$ (except for above)	6.67e-1
$-\text{NH}_2$	-3.77e-2
$-\text{NH}-$	-1.71e-1
$>\text{N}-$	-2.59e-1
$=\text{N}-$	1.83e-1
$-\text{CN}$	3.80e-1
$-\text{NO}_2$	5.84e-1
$-\text{S}-$	5.19e-2
$=\text{S}$	8.77e-1

Table A.28:  $\pi$  Group Contributions

Groups	Contribution
$-\text{CH}_3$	3.27e-1
$-\text{CH}_2-$	2.94e-1
$>\text{CH}-$	8.77e-1
$>\text{C}<$	8.38e-1
$=\text{CH}_2$	5.30e-1
$=\text{CH}-$	2.83e-1
$=\text{C}<$	-2.89e-1
$-\text{F}$	6.22e-2
$-\text{Cl}$	1.38e-1
$-\text{Br}$	6.78e-1
$-\text{I}$	1.12e+0
$-\text{OH}$ (alcohol)	-9.87e-1
$-\text{O}-$ (nonring)	-1.69e-1
$>\text{CO}$ (nonring)	-9.00e-1
$-\text{CHO}$ (aldehyde)	-7.13e-1
$-\text{COOH}$ (acid)	-7.55e-1
$-\text{COO}-$ (ester)	-4.63e-1
$-\text{NH}_2$	-8.59e-1
$-\text{NH}-$	-2.03e-1
$>\text{N}-$	-8.36e-1
$=\text{N}-$	3.73e-1
$-\text{CN}$	-5.73e-1
$-\text{NO}_2$	-3.68e-1
$-\text{S}-$	5.34e-1

# Appendix B

## Estimation Procedures

Estimation techniques are combined into estimation procedures in both the interactive and automatic design procedures. The potential error of an estimation procedure is thus greater than the error of any of the component estimation techniques. In this appendix I evaluate the accuracy of the estimation procedures used in my thesis work.

### B.1 $P_{vp}$ Estimation Procedures

Refrigerant design required estimating the vapor pressure at the evaporator and condenser temperatures. Different estimation procedures were developed for the interactive and automatic design procedures.

#### B.1.1 Interactive Estimation Procedure

The interactive design's estimation procedure used factor analytic relationships to reduce the design space to two dimensions. Riedel-Plank-Miller's  $P_{vp}$  technique, factor

relationships for  $T_b$ ,  $T_{br}$ , and  $P_c$ , and group contribution techniques for  $F_1$ ,  $F_2$ , and  $F_3$  were combined into the following estimation procedure:

- 1)  $P_{vp} = P_{vp}(T_b, T_c, P_c)$  – Riedel-Plank-Miller
- 2)  $T_b = T_b(F_1, F_2, F_3)$  –  $T_b$  Factor Model
- 2)  $T_c = T_c(F_1, F_2, F_3)$  –  $T_c$  Factor Model
- 2)  $P_c = P_c(F_1, F_2, F_3)$  –  $P_c$  Factor Model
- 3)  $F_1 = F_1(\text{groups})$  – Joback  $F_1$
- 3)  $F_2 = 0$  –  $F_2$  Assumption
- 3)  $F_3 = F_3(\text{groups})$  – Joback  $F_3$

Table B.1 shows the estimation errors for 25 compounds.  $P_{vp}$  was estimated at two temperatures for 40 compounds. The average absolute error was 5.717 bar with a standard deviation of 9.74. The average absolute percent error is 3068% with a standard deviation of 6327.

### B.1.2 Automatic Estimation Procedure

The estimation procedure for the vapor pressure used in the automatic design follows:

- 1)  $P_{vp} = P_{vp}(T_b, T_{br}, P_c)$  – Riedel-Plank-Miller EOT
- 2)  $T_b = T_b(\text{groups})$  – Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(\text{groups})$  – Joback  $T_{br}$  GCT
- 2)  $P_c = P_c(\text{groups})$  – Joback  $P_c$  GCT

Table B.2 shows the estimation errors for 25 compounds.  $P_{vp}$  was estimated for 40 compounds at two temperatures. The average absolute error is 0.221 bar with a standard deviation of 0.289. The average absolute percent error is 48.54% with a standard deviation of 48.01.

Table B.1:  $P_{vp}$  Estimation Procedure Errors – Interactive

Compound	T(K)	Literature	abs(Error)	abs(% Error)
trifluorobromomethane	214.79	0.992	0.965	97.2
trichlorofluoromethane	275.50	0.446	0.860	192.7
carbon tetrachloride	313.15	0.285	7.160	2516.3
dichloromonofluoromethane	264.39	0.402	2.230	554.7
chloroform	270.54	0.070	5.852	8399.3
methyl bromide	270.85	0.802	22.203	2767.4
methyl iodide	272.36	0.179	25.682	14330.7
nitromethane	328.86	0.199	14.095	7075.8
1,2-dichloroethane	273.22	0.028	1.916	6843.1
acetic acid	302.95	0.027	8.891	33350.2
ethyl chloride	271.27	0.576	1.836	318.6
ethanol	292.77	0.057	18.434	32192.8
dimethyl sulfide	273.25	0.224	0.210	93.5
acrylonitrile	313.15	0.280	9.073	3240.6
propionitrile	270.20	0.013	1.132	8400.6
acetone	271.76	0.087	1.536	1767.2
propionic acid	328.26	0.029	3.905	13508.1
methyl ethyl ketone	314.61	0.252	1.022	406.3
1-chlorobutane	273.95	0.039	0.012	30.1
2-butanol	322.88	0.106	1.496	1410.0
ethyl ether	273.16	0.248	0.224	90.3
diethyl sulfide	319.07	0.199	0.196	98.3
diethyl amine	304.60	0.400	0.369	92.3
1,4-pentadiene	275.60	0.406	0.336	82.7
triethylamine	323.15	0.260	0.259	99.6

Error = Estimate – Literature.

Table B.2:  $P_{vp}$  Estimation Procedure Errors – Automatic

Compound	T(K)	Literature	abs(Error)	abs(% Error)
trifluorobromomethane	214.79	0.992	0.979	98.7
trichlorofluoromethane	275.50	0.446	0.383	85.8
dichloromonofluoromethane	264.39	0.402	0.199	49.6
chloroform	270.54	0.070	0.029	41.0
methyl iodide	272.36	0.179	0.026	14.7
nitromethane	328.86	0.199	0.022	10.9
1,2-dichloroethane	273.22	0.028	0.068	244.3
acetic acid	302.95	0.027	0.011	40.4
ethyl bromide	316.71	1.208	0.023	1.9
ethanol	292.77	0.057	0.058	100.6
dimethyl sulfide	273.25	0.224	0.060	26.7
acrylonitrile	313.15	0.280	0.197	70.4
propionitrile	270.20	0.013	0.010	74.7
1,2-dichloropropane	288.15	0.044	0.026	58.9
acetone	271.76	0.087	0.001	0.9
propionic acid	328.26	0.029	0.009	30.7
isopropyl chloride	273.15	0.256	0.038	14.8
1-butyne	271.59	0.684	0.023	3.3
1,3-butadiene	262.75	0.800	0.450	56.3
methyl ethyl ketone	314.61	0.252	0.016	6.4
1-chlorobutane	273.95	0.039	0.018	45.2
2-butanol	322.88	0.106	0.050	47.4
ethyl ether	273.16	0.248	0.116	46.6
diethyl sulfide	319.07	0.199	0.019	9.5
triethylamine	323.15	0.260	0.047	18.2

Error = Estimate – Literature.

## B.2 $H_v$ Estimation Procedures

Refrigerant design required estimating the enthalpy of vaporization at the evaporator temperature. Different estimation procedures were developed for the interactive and automatic design procedures.

### B.2.1 Interactive Estimation Procedure

The estimation procedure for the enthalpy of vaporization used in the interactive design follows:

- 1)  $\Delta H_v = \Delta H_v(\Delta H_{vb}, T_b, T_c) - \text{Watson EOT}$
- 2)  $\Delta H_{vb} = \Delta H_{vb}(F_1, F_2, F_3) - \Delta H_{vb} \text{ Factor EOT}$
- 2)  $T_b = T_b(F_1, F_2, F_3) - T_b \text{ Factor EOT}$
- 2)  $T_c = T_c(F_1, F_2, F_3) - T_c \text{ Factor EOT}$ 
  - 3)  $F_1 = F_1(\text{groups}) - \text{Joback } F_2 \text{ GCT}$
  - 3)  $F_2 = 0 - F_2 \text{ assumption}$
  - 3)  $F_3 = F_3(\text{groups}) - \text{Joback } F_3 \text{ GCT}$

Table B.3 shows the estimation errors for 25 compounds. I estimated  $H_v$  for 52 compounds. The average absolute error was 13.59 kJ/g-mol with a standard deviation of 7.50 kJ/g-mol. The average absolute percent error was 46.18% with a standard deviation of 27.15%..

### B.2.2 Automatic Estimation Procedure

The estimation procedure for the enthalpy of vaporization used in the automatic design follows:

Table B.3:  $H_c$  Estimation Procedure Errors – Interactive

Compound	T(K)	Literature	abs(Error)	abs(% Error)
carbon tetrachloride	298.2	32.43	13.31	41.1
chloroform	298.2	31.34	16.72	53.3
dichloromethane	298.2	28.85	16.58	57.5
methyl bromide	276.7	23.91	14.12	59.1
nitromethane	318.3	37.17	20.96	56.4
acetonitrile	298.2	32.94	18.92	57.4
1,2-dichloroethane	298.2	35.15	15.37	43.7
acetic acid	298.2	23.36	6.45	27.6
ethyl bromide	304.6	27.55	12.33	44.8
ethane	289.0	7.04	11.00	156.2
dimethyl ether	248.3	21.51	2.31	10.7
ethanol	298.2	42.30	30.45	72.0
dimethyl sulfide	275.9	28.78	10.09	35.1
dimethylamine	278.9	26.48	0.93	3.5
acetone	300.4	30.84	10.68	34.6
propionaldehyde	286.2	30.28	7.56	25.0
propionic acid	298.2	31.14	6.83	21.9
1-propanol	298.2	47.49	28.11	59.2
methyl ethyl sulfide	301.7	31.64	13.51	42.7
n-propyl amine	298.2	31.26	1.59	5.1
trimethyl amine	276.0	22.94	13.47	58.7
1-chlorobutane	298.2	33.52	0.01	0.0
ethyl ether	280.7	28.15	8.75	31.1
diethyl sulfide	324.7	34.25	17.42	50.9
n-butyl amine	298.2	35.71	4.73	13.2

Error = Estimate – Literature.

- 1)  $\Delta H_v = \Delta H_v(\Delta H_{vb}, T_b, T_{br})$  – Watson EOT
- 2)  $\Delta H_{vb} = \Delta H_{vb}(\text{groups})$  – Joback  $\Delta H_{vb}$  GCT
- 2)  $T_b = T_b(\text{groups})$  – Joback  $T_b$  GCT
- 2)  $T_{br} = T_{br}(\text{groups})$  – Joback  $T_{br}$  GCT

Table B.4 shows the estimation errors for 25 compounds. I estimated  $H_v$  for 53 compounds. The average absolute error was 2.75 kJ/mol with a standard deviation of 3.60 kJ/mol. The average absolute percent error was 11.98% with a standard deviation of 23.95%.

## B.3 $C_{pL}$ Estimation Procedures

Refrigerant design required estimating the liquid heat capacity at the average of the evaporator and condenser temperatures. Different estimation procedures were developed for the interactive and automatic design procedures.

### B.3.1 Interactive Design Procedure

The estimation procedure for the liquid heat capacity used in the interactive design follows:

- 1)  $C_{pL} = C_{pL}(\omega, C_p^\circ, T_c)$  – Rowlinson EOT
- 2)  $\omega = \omega(T_{br}, P_c)$  – Lee-Kesler EOT
- 2)  $C_p^\circ \approx C_p^\circ(298K)$  –  $C_p$  approximation
- 3)  $T_{br} = T_{br}(T_c, T_b)$  – definition
- 4)  $C_p^\circ(298K) = C_p^\circ(F_1, F_2, F_3) - C_p^\circ \text{ Factor EOT}$
- 4)  $T_c = T_c(F_1, F_2, F_3) - T_c \text{ Factor EOT}$

Table B.4:  $H_v$  Estimation Procedure Errors – Automatic

Compound	T(K)	Literature	abs(Error)	abs(% Error)
carbon tetrachloride	298.20	32.43	5.47	16.9
chloroform	298.20	31.34	1.26	4.0
methyl bromide	276.70	23.91	0.93	3.9
nitromethane	318.30	37.17	0.42	1.1
methanol	298.20	37.43	1.63	4.4
acetaldehyde	294.20	25.71	1.05	4.1
acetic acid	298.20	23.36	19.73	84.5
ethyl bromide	304.60	27.55	0.70	2.5
ethane	289.00	7.04	10.67	151.5
dimethyl ether	248.30	21.51	0.92	4.3
propionitrile	298.20	36.12	0.82	2.3
acetone	300.40	30.84	0.53	1.7
propionaldehyde	286.20	30.28	0.57	1.9
propane	277.60	16.28	5.49	33.7
1-propanol	298.20	47.49	3.42	7.2
isopropyl alcohol	298.20	45.23	1.74	3.8
methyl ethyl sulfide	301.70	31.64	0.62	2.0
trimethyl amine	276.00	22.94	1.64	7.2
1,2-butadiene	273.30	24.62	0.54	2.2
ethyl ether	280.70	28.15	0.41	1.4
diethyl sulfide	324.70	34.25	0.86	2.5
n-butyl amine	298.20	35.71	3.78	10.6
2,2-dimethylpropane	270.30	23.45	4.08	17.4
1-pentanol	298.20	56.94	4.23	7.4
n-octane	298.20	41.48	2.35	5.7

Error = Estimate – Literature.

- 4)  $T_b = T_b(F_1, F_2, F_3) - T_b$  Factor EOT
- 4)  $P_c = P_c(F_1, F_2, F_3) - P_c$  Factor EOT
- 5)  $F_1 = F_1(\text{groups}) - \text{Joback } F_1$  GCT
- 5)  $F_2 = 0 - F_2$  assumption
- 5)  $F_3 = F_3(\text{groups}) - \text{Joback } F_3$  GCT

Table B.5 shows the estimation errors for 24 compounds. I estimated  $C_{pL}$  for 30 compounds. The average absolute error was 49.71 J/mol·K with a standard deviation of 127.10 J/mol·K. The average absolute percent error was 47.50% with a standard deviation of 160.69%. Methanol was one of the compounds. Its  $C_{pL}$  estimate of 794.5 was extremely deviant from the literature value of 80.03. Excluding this value the average absolute error for the remaining 29 compounds was 26.78 J/mol·K with a standard deviation of 20.12 J/mol·K. The average absolute percent error was 18.35% with a standard deviation of 18.63%.

### B.3.2 Automatic Design Procedure

The estimation procedure for the liquid heat capacity used in the automatic design follows:

- 1)  $C_{pL} = C_{pL}(\omega, C_p^\circ, T_c)$  by Rowlinson EOT
- 2)  $\omega = \omega(T_{br}, P_c)$  by Lee-Kesler EOT
- 2)  $T_c = T_c(T_b, T_{br})$  by definition
- 2)  $C_p^\circ = C_p^\circ(C_{p,a}^\circ, C_{p,b}^\circ, C_{p,c}^\circ, C_{p,d}^\circ)$  by  $C_p^\circ$  cubic fit
  - 3)  $T_{br} = T_{br}(\text{groups})$  by Joback  $T_{br}$  GCT
  - 3)  $T_b = T_b(\text{groups})$  by Joback  $T_b$  GCT
  - 3)  $P_c = P_c(\text{groups})$  by Joback  $P_c$  GCT

Table B.5:  $C_{pL}$  Estimation Procedure Errors – Interactive

Compound	T(K)	Literature	abs(Error)	abs(% Error)
methyl chloride	293.16	80.69	1.88	2.3
methyl mercaptan	280.00	89.03	79.91	89.8
ethyl chloride	290.00	103.31	0.18	0.2
ethane	265.25	93.86	3.00	3.2
ethanol	294.31	109.87	25.72	23.4
ethyl mercaptan	293.48	117.30	81.08	69.1
acrylonitrile	290.00	107.94	7.15	6.6
propene	299.77	99.33	14.64	14.7
acetone	296.99	124.68	14.74	11.8
propane	299.77	113.36	13.94	12.3
1-propanol	290.00	139.30	33.32	23.9
1-butene	294.22	128.69	13.94	10.8
methyl ethyl ketone	290.00	157.60	19.12	12.1
n-butane	270.00	132.41	24.98	18.9
isobutane	294.22	136.99	24.17	17.6
n-butanol	303.16	179.79	44.83	24.9
ethyl ether	290.00	170.70	0.59	0.3
1-pentene	294.22	154.37	18.11	11.7
methyl n-propyl ketone	290.00	182.70	14.26	7.8
n-pentane	280.00	161.55	25.87	16.0
1-pentanol	290.00	202.01	37.32	18.5
n-hexane	288.50	190.02	26.85	14.1
2,2-dimethyl butane	280.00	182.13	53.14	29.2
2,3-dimethyl butane	280.00	182.00	44.47	24.4
1-hexanol	290.01	232.46	37.72	16.2
diisopropyl ether	290.00	213.10	26.73	12.5
n-heptane	270.00	214.81	33.57	15.6
2-methylhexane	292.40	219.27	31.13	14.2
n-octane	290.00	250.96	24.36	9.7

Error = Estimate – Literature.

- 3)  $C_{p,a}^o = C_{p,a}^o(\text{groups})$  by Joback  $C_{p,a}^o$  GCT
- 3)  $C_{p,b}^o = C_{p,b}^o(\text{groups})$  by Joback  $C_{p,b}^o$  GCT
- 3)  $C_{p,c}^o = C_{p,c}^o(\text{groups})$  by Joback  $C_{p,c}^o$  GCT
- 3)  $C_{p,d}^o = C_{p,d}^o(\text{groups})$  by Joback  $C_{p,d}^o$  GCT

Table B.6 shows the estimation errors for 25 compounds. I estimated  $C_{p,L}$  for 30 compounds. The average absolute error was 22.05 J/mol·K a standard deviation 10.71 J/mol·K. The average absolute percent error was 14.78% with a standard deviation of 7.39%.

Table B.6:  $C_{pL}$  Estimation Procedure Errors – Automatic

Compound	T(K)	Literature	abs(Error)	abs(% Error)
methyl chloride	293.16	80.69	11.25	13.9
methanol	290.10	80.03	28.11	35.1
methyl mercaptan	280.66	89.03	1.70	1.9
ethyl chloride	290.00	103.31	16.69	16.2
ethane	265.25	93.86	2.67	2.8
ethanol	294.31	109.87	29.90	27.2
ethyl mercaptan	293.48	117.30	3.67	3.1
acrylonitrile	290.00	107.94	30.93	28.7
propene	299.77	99.33	19.48	19.6
acetone	296.99	124.68	12.64	10.1
propane	299.77	113.36	18.20	16.1
1-propanol	290.00	139.30	32.56	23.4
1-butene	294.22	128.69	17.85	13.9
methyl ethyl ketone	290.00	157.60	9.96	6.3
n-butane	270.00	132.41	20.79	15.7
isobutane	294.22	136.99	19.13	14.0
n-butanol	303.16	179.79	27.65	15.4
ethyl ether	290.00	170.70	8.60	5.0
methyl n-propyl ketone	290.00	182.70	17.33	9.5
1-pentanol	290.00	202.01	38.28	18.9
2,2-dimethyl butane	280.00	182.13	23.57	12.9
2,3-dimethyl butane	280.00	182.00	27.64	15.2
1-hexanol	290.01	232.46	43.73	18.8
n-heptane	270.00	214.81	32.19	15.0
n-octane	290.00	250.96	37.03	14.8

Error = Estimate – Literature.

# Appendix C

## Physical Property Ranges

Numerous times during the methodology it is necessary to know typical ranges for physical property values. This is especially true when determining monotonicity using derivative inspection. I examined a set of compounds which corresponds closely to those listed in the data-bank of Reid, et.al.[101]. In this appendix I present typical ranges for the important physical properties in my research.

### C.1 Critical Temperature

Table C.1 lists the critical temperature for 31 compounds. The compounds were taken from a sorted list of 453. Compounds with the 14 smallest and 17 largest  $T_c$ 's are shown.

Table C.1 shows that  $T_c$  ranges from 5.2K to 926K. The majority of the low  $T_c$  compounds are gases. I chose the lower limit of  $T_c$  to be 225K. This corresponds to a value slightly lower than that of carbon tetrafluoride.  $T_c$  generally increases with

Table C.1:  $T_c$  High and Low Sample Values

Formula	Name	$T_c$ (K)
He	Helium-4	5.2
H <sub>2</sub>	Hydrogen	33.0
D <sub>2</sub>	Deuterium	38.2
Ne	Neon	44.4
N <sub>2</sub>	Nitrogen	126.2
F <sub>2</sub>	Fluorine	144.3
Ar	Argon	150.8
O <sub>2</sub>	Oxygen	154.6
NO	Nitric Oxide	180.0
CH <sub>4</sub>	Methane	190.6
Kr	Krypton	209.4
CF <sub>4</sub>	Carbon Tetrafluoride	227.6
F <sub>3</sub> N	Nitrogen Trifluoride	234.0
F <sub>4</sub> Si	Silicon Tetrafluoride	259.0
C <sub>12</sub> H <sub>10</sub> O	Diphenyl Ether	766.0
C <sub>20</sub> H <sub>42</sub>	n-Eicosane	767.0
C <sub>20</sub> H <sub>42</sub> O	1-Eicosanol	770.0
C <sub>13</sub> H <sub>12</sub>	Diphenylmethane	770.0
C <sub>19</sub> H <sub>38</sub>	n-Tetradecylcyclopentane	772.0
C <sub>11</sub> H <sub>12</sub>	1-Methylnaphthalene	772.0
C <sub>20</sub> H <sub>40</sub>	n-Pentadecylcyclopentane	780.0
C <sub>12</sub> H <sub>10</sub>	Diphenyl	789.0
C <sub>2</sub> H <sub>6</sub> O <sub>2</sub>	Ethylene Glycol	790.0
C <sub>21</sub> H <sub>42</sub>	n-Hexadecylcyclopentane	791.0
C <sub>8</sub> H <sub>4</sub> O <sub>3</sub>	Phthalic Anhydride	810.0
I <sub>2</sub>	Iodine	819.0
C <sub>14</sub> H <sub>10</sub>	Anthracene	869.3
C <sub>14</sub> H <sub>10</sub>	Phenanthrene	873.0
C <sub>18</sub> H <sub>14</sub>	o-Terphenyl	891.0
C <sub>18</sub> H <sub>14</sub>	m-Terphenyl	924.9
C <sub>18</sub> H <sub>14</sub>	p-Terphenyl	926.0

increasing molecular weight. Although the rate of increase decreases at higher molecular weight there is no natural upper limit. I chose an upper limit of 950K. The typical range of  $T_c$  was thus chosen to be [225 950] K.

## C.2 Critical Pressure

Table C.2 lists the critical pressure for 26 compounds. The compounds were taken from a sorted list of 430. Compounds with the 12 smallest and 14 largest  $P_c$ 's are shown.

From the data shown in Table C.2, I chose the typical range of  $P_c$  to be [5.0 150.0]. This range excludes Helium-4, Deuterium Oxide, and Water. The extreme values for these observations were not considered typical of organic compounds.

## C.3 Reduced Boiling Point

Table C.3 lists the reduced boiling point for 25 compounds. The compounds were taken from a sorted list of 450. Compounds are arranged with the 14 smallest first, 6 large second, and the 5 largest  $T_{b_r}$ 's third.

The physical range of  $T_{b_r}$  is [0 1]. Table C.3 shows the typical range of  $T_{b_r}$  to be [0.558 0.818].

## C.4 Normal Boiling Point

Table C.4 lists the normal boiling point for 24 compounds. The compounds were taken from a sorted list of 478. Compounds with the 15 smallest and the 9 largest  $T_b$ 's are

Table C.2:  $P_c$  High and Low Sample Values

Formula	Name	$P_c$ (bar)
He	Helium-4	2.3
C <sub>21</sub> H <sub>42</sub>	n-Hexadecylcyclopentane	9.7
C <sub>20</sub> H <sub>40</sub>	n-Pentadecylcyclopentane	10.2
C <sub>20</sub> H <sub>42</sub>	n-Eicosane	11.1
C <sub>19</sub> H <sub>40</sub>	n-Nonadecane	11.1
C <sub>19</sub> H <sub>38</sub>	n-Tetradecylcyclopentane	11.2
C <sub>18</sub> H <sub>36</sub>	1-Octadecene	11.3
C <sub>18</sub> H <sub>38</sub>	Octadecane	12.1
C <sub>18</sub> H <sub>36</sub>	n-Tridecylcyclopentane	12.1
C <sub>20</sub> H <sub>42</sub> O	1-Eicosanol	12.2
C <sub>17</sub> H <sub>34</sub>	n-Dodecylcyclopentane	13.0
H <sub>2</sub>	Hydrogen	13.0
CH <sub>6</sub> N <sub>2</sub>	Methyl Hydrazine	82.4
HI	Hydrogen Iodide	83.1
HCl	Hydrogen Chloride	83.1
HBr	Hydrogen Bromide	85.5
CH <sub>3</sub> Br	Methyl Bromide	86.1
H <sub>2</sub> S	Hydrogen Sulfide	89.4
ClNO	Nitrosyl Chloride	91.2
NO <sub>2</sub>	Nitrogen Dioxide	101.3
Br <sub>2</sub>	Bromine	103.4
H <sub>3</sub> N	Ammonia	112.8
I <sub>2</sub>	Iodine	116.5
H <sub>4</sub> N <sub>2</sub>	Hydrazine	146.9
D <sub>2</sub> O	Deuterium Oxide	216.6
H <sub>2</sub> O	Water	220.5

Table C.3:  $T_{b_r}$  High and Low Sample Values

Formula	Name	$T_{b_r}$
I <sub>2</sub>	Iodine	0.558
HI	Hydrogen Iodide	0.560
Br <sub>2</sub>	Bromine	0.564
HBr	Hydrogen Bromide	0.569
Xe	Xenon	0.569
H <sub>2</sub> S	Hydrogen Sulfide	0.572
Kr	Krypton	0.572
Cl <sub>2</sub>	Chlorine	0.573
H <sub>2</sub> O	Water	0.576
Ar	Argon	0.578
HCl	Hydrogen Chloride	0.579
D <sub>2</sub> O	Deuterium Oxide	0.581
O <sub>2</sub>	Oxygen	0.583
CH <sub>4</sub>	Methane	0.585
C <sub>4</sub> H <sub>10</sub> O <sub>3</sub>	Diethylene Glycol	0.762
C <sub>15</sub> H <sub>30</sub>	n-Decylcyclopentane	0.763
C <sub>15</sub> H <sub>32</sub>	n-Pentadecane	0.769
C <sub>15</sub> H <sub>30</sub>	1-Pentadecene	0.769
C <sub>3</sub> H <sub>8</sub> O <sub>3</sub>	Glycerol	0.775
C <sub>16</sub> H <sub>34</sub>	Hexadecane	0.775
C <sub>21</sub> H <sub>42</sub>	n-Hexadecylcyclopentane	0.805
C <sub>17</sub> H <sub>36</sub> O	Heptadecanol	0.811
C <sub>18</sub> H <sub>38</sub> O	1-Octadecanol	0.813
C <sub>20</sub> H <sub>42</sub> O	1-Eicosanol	0.816
He	Helium-4	0.818

Table C.4:  $T_b$  High and Low Sample Values

Formula	Name	$T_b$ (K)
He	Helium-4	4.25
$H_2$	Hydrogen	20.30
$D_2$	Deuterium	23.70
Ne	Neon	27.10
$N_2$	Nitrogen	77.40
$F_2$	Fluorine	85.00
Ar	Argon	87.30
$O_2$	Oxygen	90.20
$CH_4$	Methane	111.60
Kr	Krypton	119.90
NO	Nitric Oxide	121.40
$F_3N$	Nitrogen Trifluoride	144.40
$CF_4$	Carbon Tetrafluoride	145.10
Xe	Xenon	165.00
$C_2H_4$	Ethylene	169.38
$C_{18}H_{38}O$	1-Octadecanol	608.00
$C_{14}H_{10}$	Phenanthrene	612.60
$C_{14}H_{10}$	Anthracene	613.10
$C_{20}H_{42}$	n-Eicosane	617.00
$C_{20}H_{40}$	n-Pentadecylcyclopentane	625.00
$C_{20}H_{42}O$	1-Eicosanol	629.00
$C_{21}H_{42}$	n-Hexadecylcyclopentane	637.00
$C_{18}H_{14}$	m-Terphenyl	638.00
$C_{18}H_{14}$	p-Terphenyl	649.00

shown.

Table C.4 shows the typical range of  $T_b$  for small molecular weight organic compounds to be [140 650].

# Appendix D

## Monotonicity Identification

If  $F$  is monotonic in  $x$  over the interval  $X = [\underline{X} \quad \overline{X}]$  then the interval value for  $F$  is given by:

$$[\min(F(\underline{X}), F(\overline{X})) \quad \max(F(\underline{X}), F(\overline{X}))]. \quad (\text{D.1})$$

Only the end points are needed to determine the interval extension. Since point values are used in all the calculations, excess width does not occur regardless of the function's algebraic complexity. To provide tighter bounds on the interval extensions of physical property relationships, I examined relationships used identifying major terms in the equations which were monotonic over some range of the dependent variables.

Identification of monotonicity was done primarily through derivative inspection. A similar approach was used by Sacks[109] in his BOUNDER program. The derivative of the equation or equation part being examined is checked for a consistent sign for all possible values of the dependent variables. If a consistent sign is found then the *equation or equation part is monotonic. Interestingly, examining all values of the variables is best accomplished through the use of intervals. I present the derivatives*

and results for a number of physical property relationships.

## D.1 Acentric Factor

The acentric factor is defined by

$$\omega = -\log_{10} P_{vp_r} \text{ (at } T_r = 0.7) - 1.000 \quad (\text{D.2})$$

Using the Lee-Kesler vapor pressure relationship[72] we obtain the following correlation for  $\omega$ :

$$\omega = \frac{-\ln P_c - 5.92714 + 6.09648/T_{br} + 1.28862 \ln T_{br} - 0.169347 T_{br}^6}{15.2518 - 15.6875/T_{br} - 13.4721 \ln T_{br} + 0.43577 T_{br}^6} \quad (\text{D.3})$$

Table D.1 shows the values for the acentric factor as a function of  $T_{br}$  and  $P_c$ . Figure D.1 shows this data plotted for a range of  $T_{br}$  and  $P_c$  values. Both the table and figure show that  $\omega$  is not monotonic. However, further investigation will show that the both the numerator and denominator of Equation D.3 are monotonic. I define two terms for the numerator and denominator:

$$\omega_{num} = -\ln P_c - 5.92714 + 6.09648/T_{br} + 1.28862 \ln T_{br} \quad (\text{D.4})$$

$$-0.169347 T_{br}^6$$

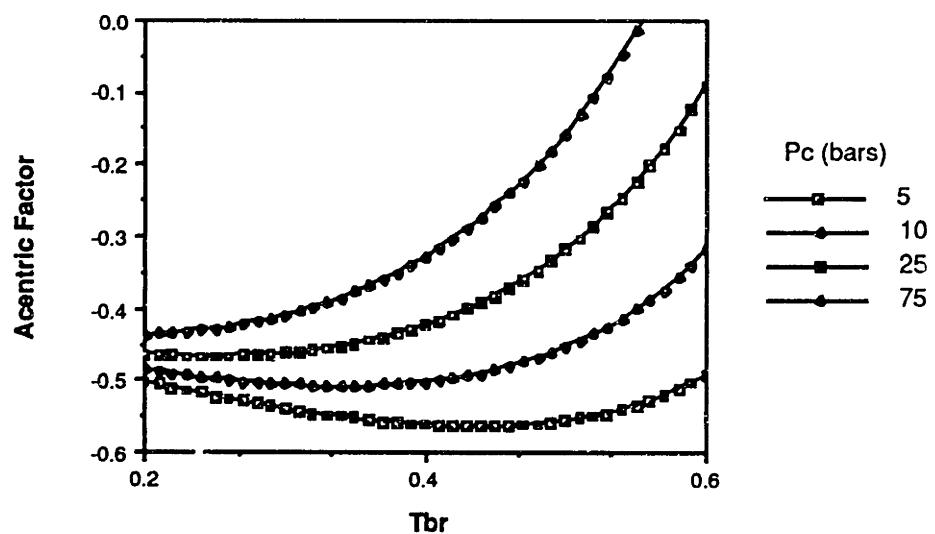
$$\omega_{den} = 15.2518 - 15.6875/T_{br} - 13.4721 \ln T_{br} + 0.43577 T_{br}^6 \quad (\text{D.5})$$

The numerator term,  $\omega_{num}$ , can be shown to be monotonic in  $T_{br}$  by analyzing its derivative. The derivative of  $\omega_{num}$  with respect to  $T_{br}$  is given by:

$$\frac{\partial \omega_{num}}{\partial T_{br}} = \frac{-6.09648}{T_{br}^2} + \frac{1.28862}{T_{br}} - 6(0.169347)T_{br}^5 \quad (\text{D.6})$$

Table D.1: Acentric Factor

$T_{br}$	$P_c = 5$	$P_c = 10$	$P_c = 25$	$P_c = 75$
0.20	-0.503	-0.487	-0.464	-0.438
0.25	-0.523	-0.499	-0.467	-0.429
0.30	-0.540	-0.507	-0.463	-0.410
0.35	-0.554	-0.509	-0.449	-0.378
0.40	-0.562	-0.503	-0.424	-0.329
0.45	-0.564	-0.486	-0.383	-0.258
0.50	-0.557	-0.454	-0.319	-0.157
0.55	-0.535	-0.402	-0.226	-0.015
0.60	-0.494	-0.320	-0.091	0.185
0.65	-0.423	-0.195	0.106	0.466
0.70	-0.307	-0.006	0.392	0.869
0.75	-0.119	0.286	0.821	1.463
0.80	0.190	0.750	1.491	2.379
0.85	0.734	1.550	2.627	3.919
0.90	1.848	3.162	4.898	6.979
0.95	5.187	7.959	11.624	16.019


 Figure D.1: Acentric Factor as a Function of  $T_{br}$  and  $P_c$

Beginning with the requirement that

$$0 < T_{b_r} < 1$$

simple algebraic manipulation yields

$$\frac{1}{T_{b_r}^2} > \frac{1}{T_{b_r}}.$$

Noting that

$$6.09648 > 1.28862$$

we conclude

$$\frac{6.09648}{T_{b_r}^2} > \frac{1.28862}{T_{b_r}}. \quad (\text{D.7})$$

Equation D.7 implies that the first term of Equation D.6, which is negative, is always greater than the second term, thus producing a negative result which is added to the negative third term. The derivative is always negative.  $\omega_{num}$  is thus monotonic.

$\omega_{den}$  is shown to be monotonic by a similar argument. The derivative of the denominator with respect to  $T_{b_r}$  is:

$$\frac{\partial \omega_{den}}{\partial T_{b_r}} = \frac{15.6875}{T_{b_r}^2} - \frac{13.4721}{T_{b_r}} + 6(0.43577)T_{b_r}^5. \quad (\text{D.8})$$

The first term of Equation D.8, which is positive, is always greater than the second term. Their difference is then always positive which is added to the positive third term. The derivative is always positive.  $\omega_{den}$  is thus monotonic.

The following sample calculation shows the effect taking advantage of monotonicity has on the dependency width of an interval. I determine the value of the acentric factor given

$$T_{b_r} = [0.7 \quad 0.9]K \quad \text{and} \quad P_c = [25 \quad 40]atm.$$

First calculating  $\omega_{num}$  and  $\omega_{den}$  using their simple interval extensions we obtain:

$$\omega_{num} = [-3.392 \quad -0.592] \quad (D.9)$$

$$\omega_{den} = [-5.688 \quad 2.858]. \quad (D.10)$$

Taking advantage of the monotonicity of  $\omega_{num}$  and  $\omega_{den}$  we can compute the tighter intervals:

$$\omega_{num} = [-3.068 \quad -0.916] \quad (D.11)$$

$$\omega_{den} = [-2.302 \quad -0.528]. \quad (D.12)$$

Besides the reduced width of the monotonically derived intervals as compared to the naively evaluated intervals it is important to note that the simple interval extension of  $\omega_{den}$  gives an interval which contains zero. Since  $\omega_{den}$  is the denominator of Equation D.3, the interval value we get for  $\omega$  is  $[-\infty \quad \infty]$ . Using tighter intervals values the acentric factor is computed to be  $[0.602 \quad 5.811]$ .

## D.2 Vapor Pressure

I spent considerable time investigating the monotonicity of the Riedel-Plank-Miller vapor pressure estimation technique. The estimate for  $P_{vp}$  is given by four equations:

$$\ln P_{vp_r} = -\frac{G}{T_r} \left[ 1 - T_r^2 + k (3 + T_r) (1 - T_r)^3 \right] \quad (D.13)$$

$$k = \frac{h/G - (1 + T_{br})}{(3 + T_{br}) (1 - T_{br})^2} \quad (D.14)$$

$$G = 0.4835 + 0.4605h \quad (D.15)$$

$$h = T_{br} \frac{\ln P_c}{1 - T_{br}} \quad (D.16)$$

The independent variables in the equations were taken to be  $T_{br}$ ,  $P_c$ , and  $T_b$ . A typical set of values such as:

$$T_{br} = [0.7 \quad 0.8] \quad P_c = [25 \quad 50] \quad T_b = [300 \quad 500]$$

gives

$$P_{vp} = [0 \quad \infty] \text{ at } 272.05\text{K}.$$

If we assume  $P_{vp}$  is monotonic in these properties then the resulting interval is:

$$P_{vp} = [0 \quad 0.311] \text{ at } 272.05\text{K}.$$

The large reduction in interval width indicates that investigating monotonicity is worthwhile.

I used the derivative inspection method to investigate monotonicity. Due to the algebraic complexity of Eqs. D.13 through D.16 monotonicity was sometimes investigated for only part of the equations. I discuss a monotonicity investigation in each of the following sections.

### D.2.1 $P_{vp}$ Monotonicity

Inspection of  $P_{vp}$ 's derivatives with respect to  $T_b$  and  $P_c$  identified monotonicity. Inspection of  $P_{vp}$ 's derivative with respect to  $T_{br}$  yielded inconclusive results. Subexpressions were analyzed for monotonicity with respect to  $T_{br}$ .

The derivative of  $P_{vp}$  with respect to  $T_b$  is given by:

$$\frac{\partial P_{vp}}{\partial T_b} = P_c e^\sigma G \frac{\partial T_r}{\partial T_b} \left[ \frac{1}{T_r^2} + 1 + \frac{3k}{T_r^2} (1 - T_r)^2 (1 + T_r)^2 \right] \quad (\text{D.17})$$

$$\alpha = -\frac{G}{T_r} \left[ 1 - T_r^2 + k(3 + T_r)(1 - T_r)^3 \right] \quad (\text{D.18})$$

$$\frac{\partial T_r}{\partial T_b} = -\frac{T T_{b_r}}{T_b^2} \quad (\text{D.19})$$

Equation D.17 shows the sign of  $\frac{\partial P_{vp}}{\partial T_b}$  is determined by the sign of the term:

$$\text{Term} = 1 + T_r^2 + 3k(1 - T_r)^2(1 + T_r)^2. \quad (\text{D.20})$$

I investigated this term for sign consistency.

I divided interval values for  $T_b$ ,  $T_{b_r}$ , and  $P_c$  into subintervals and used a united extension to evaluate *term*. The interval for  $T_b$ , [140 650], was divided into a total of 20 subintervals. The interval for  $T_{b_r}$ , [0.5 0.9], was divided into a total of 20 intervals. The interval for  $P_c$ , [5 150], was divided into 10 subintervals. The united extension of the investigated expression is [0.36 189.4] at 272.05K and [0.35 82.0] at 316.4K.  $P_{vp}$  is thus monotonic with respect to  $T_b$ .

The derivative of  $P_{vp}$  with respect to  $P_c$  is given by:

$$\begin{aligned} \frac{\partial P_{vp}}{\partial P_c} &= e^\alpha \left[ 1 - \frac{P_c}{T_r} \left( 1 - T_r^2 + \frac{1.05}{G} \frac{(3 + T_r)(1 - T_r)^3}{(3 + T_{b_r})(1 - T_{b_r})^2} \right. \right. \\ &\quad \left. \left. + k(3 + T_r)(1 - T_r)^3 \right) \frac{\partial g}{\partial P_c} \right] \end{aligned} \quad (\text{D.21})$$

$$\alpha = -\frac{G}{T_r} \left[ 1 - T_r^2 + k(3 + T_r)(1 - T_r)^3 \right] \quad (\text{D.22})$$

$$\frac{\partial g}{\partial P_c} = \frac{0.4605}{P_c} \frac{T_{b_r}}{(1 - T_{b_r})} \quad (\text{D.23})$$

The sign of  $\frac{\partial P_{vp}}{\partial P_c}$  is determined by:

$$\text{Term} = 1 - \frac{P_c}{T_r} \left( 1 - T_r^2 + \frac{1.05}{G} \frac{(3 + T_r)(1 - T_r)^3}{(3 + T_{b_r})(1 - T_{b_r})^2} + k(3 + T_r)(1 - T_r)^3 \right) \frac{\partial g}{\partial P_c} \quad (\text{D.24})$$

The exponential term in Equation D.21 does not affect the sign of  $\frac{\partial P_{vp}}{\partial P_c}$ .

I originally determined that  $P_{vp}$  was monotonic with respect to  $P_c$ . Repeating the investigation I question that finding. The estimation technique used for  $P_{vp}$  did assume monotonicity in  $P_c$ .

### D.2.2 $h$ Monotonicity

Equation D.16 shows  $h$  is a function of  $T_{br}$  and  $P_c$ .  $T_{br}$  and  $P_c$  are always positive.  $T_{br}$  ranges from 0 to 1. Since  $P_c$  occurs only once in Equation D.16 monotonicity with respect to  $P_c$  is not investigated.  $T_{br}$  occurs twice. The derivative of  $h$  with respect to  $T_{br}$  is given by:

$$\frac{\partial h}{\partial T_{br}} = \frac{\ln P_c}{(1 - T_{br})^2}. \quad (\text{D.25})$$

For all practical values of  $P_c$  and  $T_{br}$ ,  $\frac{\partial h}{\partial T_{br}}$  is positive. Therefore  $h$  is monotonic with respect to  $T_{br}$ .

### D.2.3 $G$ Monotonicity

Equation D.15 shows  $G$  is a function of  $h$  and thus indirectly a function of  $T_{br}$  and  $P_c$ . I do not investigate monotonicity with respect to  $P_c$  since there is only a single occurrence. The derivative of  $G$  with respect to  $T_{br}$  equals:

$$\frac{\partial g}{\partial T_{br}} = 0.4605 \frac{\partial h}{\partial T_{br}}. \quad (\text{D.26})$$

Section D.2.2 showed  $\frac{\partial h}{\partial T_{br}}$  was positive for all practical values of  $P_c$  and  $T_{br}$ . Therefore  $\frac{\partial G}{\partial T_{br}}$  is always positive.  $G$  is thus monotonic with respect to  $T_{br}$ .

#### D.2.4 $k$ Monotonicity

Equation D.14 shows  $k$  is a function of  $h$ ,  $G$ , and  $T_{b_r}$ .  $k$  is thus indirectly a function of  $T_{b_r}$  and  $P_c$ . The derivative of  $k$  with respect to  $T_{b_r}$  is given by:

$$\frac{\partial k}{\partial T_{b_r}} = \frac{(3 + T_{b_r})(1 - T_{b_r}) \left[ \frac{\partial h/G}{\partial T_{b_r}} - 1 \right] + \left[ \frac{h}{G} - (1 + T_{b_r}) \right] (5 + 3T_{b_r})}{(3 + T_{b_r})^2 (1 - T_{b_r})^3} \quad (\text{D.27})$$

$$\frac{\partial h/G}{\partial T_{b_r}} = \frac{0.4835 \ln P_c}{G^2 (1 - T_{b_r})^2} \quad (\text{D.28})$$

I divided interval values for  $T_{b_r}$  and  $P_c$  into subintervals and used a united extension to evaluate  $\frac{\partial k}{\partial T_{b_r}}$ . The interval for  $T_{b_r}$ ,  $[0.1 \quad 0.9]$ , was divided into 1000 subintervals. The interval for  $P_c$ ,  $[5 \quad 150]$ , was divided into 200 subintervals. The united extension of  $\frac{\partial k}{\partial T_{b_r}}$  is  $[0.11 \quad 102.2]$ .  $k$  is thus monotonic with respect to  $T_{b_r}$ .

# Appendix E

## Factor Analysis

Factor analysis and principal component analysis are two statistical techniques for analyzing covariance structure. The goal of each technique is to explain, if possible, the covariance relationships among many variables in terms of a few underlying quantities. These new quantities are called *factors* in factor analysis or *principal components* in principal component analysis. Factors and principal components are new variables formed from linear combinations of the original observed variables.

In my molecular design procedure factor analysis plays an important role in being able to reduce the dimensionality of a design space. Interactive design in a space of dimension greater than three is not possible. Interactive design in a three dimensional space is possible but difficult. It is therefore desirable to reduce the dimension of the space to two.

I present in this appendix both statistical techniques of factor analysis and principal component analysis. I first present the mathematical basis of the techniques. I then describe some of the work done in analyzing physical properties using factor analysis.

I do not describe the two techniques in detail. Both are common statistical techniques described in many books on multivariate statistical analysis[63,82].

## E.1 Principal Component Analysis

Principal component analysis looks for a *few* linear combinations which can be used to summarize a set of data, losing in the process as little information as possible. The definition of principal components follows[82]:

If  $x$  is a random vector with mean  $\mu$  and covariance matrix  $\Sigma$ , then the principal component transformation is the transformation:

$$x \rightarrow y = \Gamma'(x - \mu), \quad (\text{E.1})$$

where  $\Gamma$  is orthogonal,  $\Gamma'\Sigma\Gamma = \Lambda$  is diagonal, and  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ .

The strict positivity of the eigenvalues,  $\lambda_i$ , is guaranteed if  $\Sigma$  is positive definite. This representation of  $\Sigma$  follows from the spectral decomposition theorem. The  $i$ th principal component of  $x$  may be defined as the  $i$ th element of the vector  $y$ :

$$y_i = \gamma'_{(i)}(x - \mu). \quad (\text{E.2})$$

Here  $\gamma_{(i)}$  is the  $i$ th column of  $\Gamma$ .

Geometrically, principal components represent the selection of a new coordinate system obtained by rotating the original system with  $X_1, X_2, \dots, X_n$  as the coordinate axes. The new axes,  $Y_1, Y_2, \dots, Y_n$ , represent the directions with maximum variability.

## E.2 Factor Analysis

Factor analysis is similar to principal component analysis. The major difference is that factor analysis attempts separate the variation in a variable into a common and specific part. The definition of factor analysis follows[82]:

Let  $x(p \times 1)$  be a random vector with mean  $\mu$  and covariance matrix  $\Sigma$ .

The  $k$ -factor model holds for  $x$  if  $x$  can be written in the form

$$x = \Lambda f + u + \mu \quad (\text{E.3})$$

where  $\Lambda(p \times k)$  is a matrix of constants and  $f(k \times 1)$  and  $u(p \times 1)$  are random vectors. The elements of  $f$  are called *common factors*. The elements of  $u$  are called *specific* or *unique* factors. Additional restrictions on the model follow:

$$E(f) = 0 \quad (\text{E.4})$$

$$V(f) = I \quad (\text{E.5})$$

$$E(u) = 0 \quad (\text{E.6})$$

$$C(u_i, u_j) = 0, \quad i \neq j \quad (\text{E.7})$$

$$C(f, u) = 0. \quad (\text{E.8})$$

## E.3 Factor Analytic Studies

Factor analysis is becoming widely used in the development of physical property – molecular structure relationships. In the chemical field, Malinowski and Howery[81]

used factor analysis to analyze mass spectroscopy data, chromatography data, and much more. Numerous books on drug design describe its applicability[39,84]. I am primarily interested in the dimensional reduction which is accomplished in factor analytic studies. Three studies have looked at the interrelationships between physical properties which could be of value to my molecular design technique.

### **E.3.1 Cramer**

Cramer[20] used factor analysis to test the hypothesis that intermolecular interactions in the liquid state were of two types. The first type of interaction was associated with the “bulk” of the molecules involved. The second type of interaction was associated with the “cohesiveness” of the molecules involved.

Cramer examined 114 pure component liquids in his factor analytic study. Several combinations of properties were examined in the study. The results of a six property – five factor analysis were used to develop regression equations for 15 other properties. The factor relationships for the original six properties and the additional fifteen properties are shown in Table E.1. Summary statistics for these equations are shown in Table E.2. For many of the 21 properties the statistics show good fits.

Table E.1: Physical Properties as Functions of BC(DEF) Factors

---

1	Activity Coefficient: $\log(c_{H_2O}/c_{gas})$	$= 1.241 + 5.09B + 13.54C + 3.36D + 6.83E + 7.39F$
2	log(Partition Coefficient): $(c_{octanol}/c_{H_2O})$	$= 1.604 + 3.65B - 7.66C - 5.74D - 0.31E + 5.09F$
3	Molar Refractivity	$= 22.94 + 52.89B - 21.58C + 4.21D + 83.6E - 12.02F$
4	Boiling Point	$= 66.39 + 532.50B + 223.6C - 365.4D - 250.8E - 794.6F$
5	Molar Volume	$= 90.02 + 139.9B - 90.12C + 245.6D - 108.5E - 3.1F$
6	Heat of Vaporization	$= 8.197 + 14.92B + 9.615C - 8.07D - 12.8E + 14.5F$
7	Magnetic Susceptibility	$= 54.58 + 111.4B - 51.8C - 16.1D + 42.9E + 13.0F$
8	Critical Temperature	$= 248.6 + 770.4B + 343.1C - 927D - 68E - 1810F$
9	(van der Waals A) <sup>1/2</sup>	$= 4.144 + 6.93B - 0.23C - 0.03D + 1.2E - 10.6F$
10	van der Walls B	$= 0.121 + 0.225B - 0.07C + 0.15D + 0.11E - 0.12F$
11	log(Dielectric Constant)	$= 0.728 + 0.57B + 2.60C - 1.97D - 5.1E - 2.3F$
12	solubility parameter	$= 8.97 + 5.22B + 12.1C - 15.0D - 12.9E - 7.7F$

---

Table E.1 Continued: Physical Properties as Functions of BC(DEF) Factors

13	critical pressure	$= 46.3 - 7.1B + 56.2C - 243D - 51E - 230F$
14	surface tension	$= 23.5 + 36.8B + 22.2C - 101D + 58E - 174F$
15	thermal conductivity	$= 3.17 + 3.19B + 4.5C + 2.1D - 13.2E - 14.5F$
16	$\log(\text{viscosity})$	$= -0.23 + 2.44B + 1.26C - 0.94D - 8.1E + 5.06F$
17	isothermal compressibility	$= 11.6 - 22.6B - 10.8C + 57.6D + 11E + 22F$
18	$E_T$	$= 38.4 + 9.4B + 47.9C - 41.7D - 147E + 63F$
19	Dipole Moment	$= 1.33 + 1.5B + 5.0C - 1.5D - 10E - 20.4F$
20	Melting Point	$= 77.2 + 295B + 194C - 443D + 339E - 406F$
21	Molecular Weight	$= 87.8 + 149.6B - 78.3C - 239D + 20E - 386F$

### E.3.2 Klincewicz

Klincewicz[67] performed a factor analytic study on the thirteen variables shown in Table E.3. Studies were done using 2, 3, and 4 factor models. The percentage variance explained by each factor in each study is shown in Table E.4. The contribution of the fourth factor to the total variance is not substantial. The three factor model is considered significant.

The factor loadings for the three factor model are displayed in Table E.6. To emphasize patterns within the factors, Klincewicz replaced loadings with absolute values less

Table E.2: Statistics for BC(DEF) Factors - Physical Property Relationships

	Property	Units	$r^2$	Std Error
1	Activity Coefficient $\log(c_{H_2O}/c_{gas})$	–	0.998	0.14
2	$\log$ (Partition Coefficient) $(c_{octanol}/c_{H_2O})$	–	0.998	0.08
3	Molar Refractivity	$\text{cm}^3/\text{mol}$	0.999	0.51
4	Boiling Point	$^{\circ}\text{C}$	0.999	3.31
5	Molar Volume	$\text{cm}^3/\text{mol}$	0.999	1.27
6	Heat of Vaporization	$\text{kcal}/\text{mol}$	0.997	0.22
7	Magnetic Susceptibility	$\text{cgs molar}$	0.963	6.31
8	Critical Temperature	$^{\circ}\text{C}$	0.992	20.71
9	(van der Waals A) $^{1/2}$	$\text{l}\cdot\text{atm}^{1/2}/\text{mol}$	0.991	0.19
10	van der Walls B	$\text{l}/\text{mol}$	0.975	0.01
11	$\log$ (Dielectric Constant)	–	0.883	0.23
12	solubility parameter	$\text{cal}/\text{cm}^3$	0.912	0.80
13	critical pressure	atm	0.769	9.75
14	surface tension	$\text{dyn}/\text{cm}$	0.953	2.55
15	thermal conductivity	$\text{cal}/\text{sec}\cdot\text{cm}^2$	0.808	0.43
16	$\log$ (viscosity)	–	0.920	0.19
17	isothermal compressibility	$\text{m}^2/\text{mol}\times 10^{10}$	0.935	1.19
18	$E_T$		0.980	1.73
19	Dipole Moment	d	0.733	0.77
20	Melting Point	$^{\circ}\text{C}$	0.851	39.6
21	Molecular Weight	$\text{g}/\text{mol}$	0.779	25.8

Table E.3: Physical Properties in Klincewicz's Factor Analytic Study

$MR$	$Mw$	$\rho$	$\delta_m$
$T_c$	$P_c$	$V_c$	$T_b$
$\Delta G_{f,298}^{\circ}$	$\Delta H_{f,298}^{\circ}$	$n_A$	$C_{p,298}^{\circ}$
	$T_m$		

Table E.4: Percentage Variance Explained by Klincewicz's Factors

Factor	Number of Factors		
	2	3	4
1	76.65	69.07	70.74
2	16.33	14.37	14.22
3		14.20	12.29
4			1.84
Total	92.98	97.64	99.09

than 0.25 with zero. The means and standard deviations for the observations on the thirteen variables are shown in Table E.5. From this information it should be possible to generate physical property - factor relations using the equation:

$$Property = STD \times (L_{F_1} F_1 + L_{F_2} F_2 + L_{F_3} F_3) + Mean \quad (E.9)$$

where:

*STD* = standard deviation of the data on the physical property.

*L<sub>i</sub>* = loading of the *i*th factor.

*Mean* = average value of the data on the physical property.

The relationships I developed did not accurately reproduce the study's data. I believe the substituted zeros caused the large errors.

### E.3.3 Joback

Joback[61] performed a factor analytic study on the nine physical properties shown in Table E.7. Two, three, and four factor models were investigated for 98 compounds. The total percentage variance explained by each of these models is shown in Table E.8.

Table E.5: Klincewicz's Factor Analysis Data Summary Statistics

Physical Property	Mean	Standard Deviation
$Mw$	88.407	93.719
$T_c$	546.726	553.496
$T_b$	360.049	365.367
$T_m$	201.927	210.669
$V_c$	302.200	317.556
$\rho$	0.862	0.894
$MR$	26.420	28.225
$1/P_c$	0.252	0.260
$C_{p,298}^o$	26.563	28.203
$\Delta H_{f,298}^o$	-29.636	46.394
$\Delta G_{f,298}^o$	-3.375	33.299
$\delta_m$	0.351	0.473
$n_A$	15.027	16.141

Table E.6: Loadings for Klincewicz's Three Factor Model

Physical Property	Loadings		
	Factor 1	Factor 2	Factor 3
$V_c$	0.970	0	0
$n_A$	0.969	0	0
$C_{p,298}^o$	0.968	0	0
$MR$	0.967	0	0
$1/P_c$	0.950	0	0
$Mw$	0.929	0.319	0
$T_b$	0.904	0.402	0
$T_c$	0.898	0.416	0
$T_m$	0.849	0.460	0
$\rho$	0.825	0.493	0
$\delta_m$	0.410	0.858	0
$\Delta G_{f,298}^o$	0	0	0.975
$\Delta H_{f,298}^o$	-0.496	0	0.857

To emphasize patterns within the factors, Klincewicz replaced loadings with absolute values less than 0.25 with zeros.

---

Table E.7: Physical Properties in Joback's Factor Analytic Study

---

$V_c$	$1/\sqrt{P_c}$	$C_{p,298}^o$
$n_A$	$\Delta G_{f,298}^o$	$\Delta H_{f,298}^o$
$T_b$	$T_c$	$\Delta H_{vb}$

---

The three factor model was considered the best. Table E.9 shows the nine equations relating the factors to the original physical properties[61].

## E.4 Dimensional Reduction

It is desirable to design in a two dimensional design space when using the interactive design procedure. Many of the equation oriented estimation techniques used in target transformation yield transformed constraints which are dependent on more than two fundamental properties. The results of the factor analytic studies show that factor analysis is a powerful procedure for reducing the dimensionality of physical property data.

It is especially interesting to note the variance explained by two factors in the above studies. Table E.10 shows the variance explained by two factors in each of these studies.

---

Table E.8: Total Variance Explained in Joback's Factor Models

---

Number of Factors	% Variance Explained
2	74.78
3	96.15
4	97.24

---

Table E.9: Physical Property - Factor Relationships

$1/\sqrt{P_c}$	=	0.157	-	0.019 $F_1$	+	0.001 $F_2$		
$V_c$	=	296.1	-	89.66 $F_1$	+	10.41 $F_2$	-	36.68 $F_3$
$n_A$	=	14.50	-	5.35 $F_1$	-	0.27 $F_2$	-	1.18 $F_3$
$C_{p,298}^\circ$	=	25.70	-	8.72 $F_1$	-	0.29 $F_2$	-	2.44 $F_3$
$\Delta G_{f,298}^\circ$	=	1.41	-	6.93 $F_1$	+	31.80 $F_2$	-	0.51 $F_3$
$\Delta H_{f,298}^\circ$	=	-22.64	+	5.42 $F_1$	+	36.94 $F_2$	+	0.50 $F_3$
$T_c$	=	545.9	-	24.65 $F_1$	+	11.99 $F_2$	-	87.92 $F_3$
$T_b$	=	358.4	-	25.26 $F_1$	+	1.20 $F_2$	-	64.94 $F_3$
$\Delta H_{vb}$	=	7686.6	-	432.3 $F_1$	-	255.2 $F_2$	-	1614.3 $F_3$

Cramer reported results for factorizations of four, six, and ten physical property data sets. Each of these is shown in Table E.10. Even though the variance explained reported by Joback is low, all studies shown that two factors explain a considerable amount of the physical property data.

As discussed in Section 16 the factor analytic studies performed used a data set consisting of a wide variety of compounds. Targeting the data set for a particular chemical product should produce results of greater accuracy. In the refrigerant case study presented in Chapter 11 the factor analytic relationships used to reduce the dimensionality of the design space should be derived using data from current refrigerants.

## E.5 Group Contributions

Cramer[21] and Joback[61] developed group contribution estimation techniques for the factor scores derived in their studies. Cramer reported good accuracy for his estimation

Table E.10: Percentage Variance Explained by 2 Factors

Study	Factors		
	1	2	Total
<b>Cramer<sup>a</sup></b>			
4 Properties	55.8	41.7	97.5
6 Properties	64.4	31.3	95.7
10 Properties	75.3	21.4	96.7
Klincewicz	76.7	16.3	93.0
Joback <sup>b</sup>	—	—	74.78

<sup>a</sup> The percentage variance explained was taken from the first two factors of a five factor study (Four factors in the 4 property study).

<sup>b</sup> Only the total variance was reported.

techniques with an rms difference between predicted and experimental values for 749 compounds being 6%. Joback's estimation techniques were not as accurate.

## **Abstract**

This volume describes the computer implementation of some of the methodologies presented in Volume 1. Implementation was done on a Symbolics LISP Machine using the Genera development environment. I used flavors, commands, constraint frames, generic functions, and presentations to implement my molecular design environment.

The sections of the methodology described in Volume 1 were incorporated into separate sections of the implementation. I devote a chapter to each of these sections. In describing each section I emphasize the screen layout, objects created, important instance variables, generic functions, and commands.

The purpose of this volume is to describe my system in sufficient detail to allow easy use of the system, provide information for maintenance, and provide guidance in extending the system. Throughout the volume I discuss the concepts learned about organizing and implementing a large complex computer environment. I believe this makes the volume useful to individuals who do not know LISP, have not programmed a LISP machine, and may never have the opportunity to use the system.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Symbolics Environment . . . . .	1
1.2	System Organization . . . . .	1
1.3	Objects . . . . .	5
1.4	Commands . . . . .	6
1.5	Structure . . . . .	6
<b>2</b>	<b>System Basics</b>	<b>7</b>
2.1	Keyboard . . . . .	7
2.2	Mouse . . . . .	8
2.3	Non-Displayed Commands . . . . .	9
2.3.1	Symbolics' Commands . . . . .	12
<b>3</b>	<b>Login Section</b>	<b>13</b>
3.1	Section Layout . . . . .	13
3.2	Section Operation . . . . .	15
3.3	Section Objects . . . . .	16
3.4	Section Commands . . . . .	19
3.5	Section Discussion . . . . .	20
3.6	Example Usage . . . . .	21
<b>4</b>	<b>Problem Formulation Section</b>	<b>22</b>
4.1	Section Layout . . . . .	22
4.2	Section Operation . . . . .	24

4.2.1	Creating Constraints . . . . .	24
4.2.2	Creating Properties . . . . .	26
4.2.3	Creating Property Classes . . . . .	27
4.3	Section Objects . . . . .	27
4.3.1	Property Object . . . . .	28
4.3.2	Property Class Object . . . . .	29
4.3.3	Constraint Object . . . . .	31
4.4	Section Commands . . . . .	32
4.5	Section Discussion . . . . .	36
4.5.1	Knowledge Representation . . . . .	36
4.5.2	Property Symbols . . . . .	36
4.6	Example Usage . . . . .	37
<b>5</b>	<b>Data Base Section</b>	<b>47</b>
5.1	Data Configuration Layout . . . . .	48
5.2	Plot Configuration Layout . . . . .	50
5.3	Section Operation . . . . .	52
5.4	Data Configuration Objects . . . . .	54
5.4.1	Data Base Molecule Object . . . . .	54
5.4.2	Data Object . . . . .	56
5.4.3	Data Base Keyword . . . . .	57
5.4.4	Keyword Function Object . . . . .	58
5.4.5	Data Display Object . . . . .	59
5.4.6	Correlation Matrix Object . . . . .	61
5.5	Data Configuration Commands . . . . .	61
5.6	Plot Configuration Objects . . . . .	65
5.6.1	2D Active Graph Object . . . . .	66
5.6.2	2D Data Point Object . . . . .	68
5.6.3	Function Graphic Object . . . . .	68
5.6.4	Histogram Object . . . . .	68

5.7	Plot Configuration Commands . . . . .	69
5.8	Section Discussion . . . . .	76
5.9	Example Usage . . . . .	77
<b>6</b>	<b>Group Contribution Section</b>	<b>88</b>
6.1	Editing Configuration Layout . . . . .	88
6.2	Model Entry Configuration Layout . . . . .	91
6.3	Section Operation . . . . .	93
6.3.1	Creating Estimation Techniques . . . . .	93
6.3.2	Creating New Groups . . . . .	94
6.4	Editing Configuration Objects . . . . .	96
6.4.1	Groups . . . . .	98
6.5	Model Entry Configuration Objects . . . . .	100
6.6	Editing Configuration Commands . . . . .	102
6.7	Model Entry Configuration Commands . . . . .	104
6.8	Section Discussion . . . . .	109
6.8.1	Estimation Techniques . . . . .	110
6.8.2	Molecule Representation . . . . .	111
6.9	Example Usage . . . . .	111
<b>7</b>	<b>Target Transformation Section</b>	<b>135</b>
7.1	Section Layout . . . . .	135
7.2	Target Transformation Objects . . . . .	137
7.3	Target Transformation Commands . . . . .	141
7.4	Section Operation . . . . .	146
7.5	Section Discussion . . . . .	151
7.6	Example Usage . . . . .	155
<b>8</b>	<b>Automatic Design Section</b>	<b>162</b>
8.1	Section Layout . . . . .	162
8.2	Section Operation . . . . .	168

8.3	Meta-Groups Configuration Objects . . . . .	169
8.4	Meta-Molecules Configuration Objects . . . . .	173
8.5	Meta-Groups Configuration Commands . . . . .	173
8.6	Meta-Molecules Configuration Commands . . . . .	177
8.7	Section Discussion . . . . .	186
8.7.1	Physical Property Pruning . . . . .	186
8.7.2	Meta-Molecule Inheritance . . . . .	187
8.8	Example Usage . . . . .	188
<b>9</b>	<b>Interactive Design Section</b>	<b>229</b>
9.1	Section Layout . . . . .	230
9.2	Section Operation . . . . .	234
9.3	Preparation Configuration Objects . . . . .	239
9.4	Preparation Configuration Commands . . . . .	242
9.5	Design Configuration Commands . . . . .	244
9.6	Section Discussion . . . . .	248
9.7	Example Usage . . . . .	249
<b>10</b>	<b>Molecule Evaluation Section</b>	<b>259</b>
10.1	Section Layout . . . . .	260
10.2	Section Operation . . . . .	264
10.2.1	Estimation Procedure Development . . . . .	264
10.2.2	Physical Property Estimation . . . . .	266
10.3	Specifications Configuration Objects . . . . .	269
10.4	Values Configuration Objects . . . . .	272
10.5	Specifications Configuration Commands . . . . .	273
10.6	Values Configuration Commands . . . . .	276
10.7	Section Discussion . . . . .	278
10.8	Section Example . . . . .	281
<b>11</b>	<b>Final Remarks</b>	<b>293</b>

11.1 System Status . . . . .	293
11.2 Extending the System . . . . .	293
11.3 LISP . . . . .	294
11.4 Persistence . . . . .	295
<b>A Linear Names</b>	<b>299</b>
<b>B Preparation</b>	<b>302</b>
B.1 References . . . . .	303
B.2 Background . . . . .	304
B.3 Data Types . . . . .	304
B.4 Editor . . . . .	305
B.5 Basics . . . . .	305
B.6 Macros . . . . .	306
B.7 Debugger . . . . .	307
B.8 I/O . . . . .	307
B.9 Flavors . . . . .	308
B.10 Windows . . . . .	309
B.11 Presentations . . . . .	309
B.12 Interface Programming . . . . .	309
B.13 Proceeding . . . . .	310
<b>C Installation</b>	<b>311</b>
C.1 System Requirements . . . . .	311
C.2 System Restoration . . . . .	311
C.3 Loading . . . . .	313

# List of Figures

1.1	Example Configuration: Plot Configuration . . . . .	3
3.1	Login Section Screen . . . . .	14
3.2	Login Section Documentation Facility . . . . .	17
4.1	Problem Formulation Section Screen . . . . .	23
5.1	Data Base Section Data Configuration Screen . . . . .	49
5.2	Data Base Section Plot Configuration Screen . . . . .	51
5.3	Display of Physical Property Data in Tabular Format . . . . .	60
5.4	Correlation Matrix Display . . . . .	62
5.5	Plot Configuration with 2D Active Graph . . . . .	67
5.6	Acentric Factor Histogram . . . . .	70
5.7	Describing Several Data Points Displayed in a 2D Active Graph . . . . .	72
5.8	Histogram Statistics Display . . . . .	73
5.9	Polynomial regressions. Using marked points to exclude data. . . . .	75
6.1	Group Contribution Editing Configuration Screen . . . . .	89
6.2	Group Contribution Model Entry Configuration Screen . . . . .	92
6.3	Entry of a Group Contribution Estimation Technique . . . . .	95
6.4	Input Atoms for the Construction of a New Group . . . . .	117
6.5	Input Atoms Rearranged into the Shape of the New Group . . . . .	118
6.6	Single Bond Connections . . . . .	120
6.7	Ring Single Bond Connections . . . . .	121
6.8	Completely Connected Input Atoms . . . . .	123

6.9	Input Atoms Collected into a New Group . . . . .	124
6.10	Estimation Technique Development for Diamagnetic Susceptibility . . . . .	130
6.11	Editing an Estimation Model . . . . .	132
7.1	Target Transformation Section Screen . . . . .	136
7.2	Estimation Technique Documentation . . . . .	144
7.3	Initial State of Transformed Constraints . . . . .	147
7.4	Applicable Estimation Techniques for the <i>Pvp High</i> Transformed Constraint . . . . .	149
7.5	Applying the Riedel Plank Miller Technique to <i>Pvp High</i> . . . . .	150
7.6	Displaying Applicable Estimation Techniques . . . . .	152
7.7	Applying a Group Contribution Technique . . . . .	153
7.8	Complete Transformation to Factor Dependency . . . . .	154
8.1	Meta-Groups Configuration Screen . . . . .	163
8.2	Meta-Group 1 Expanded by Global Valence . . . . .	165
8.3	Meta-Molecules Configuration Screen . . . . .	166
8.4	Initial Set of Meta-Molecules . . . . .	170
8.5	Expansion by the Global Valence Structural Characteristic . . . . .	171
8.6	Meta-Group Statistics Display . . . . .	178
8.7	Property Constraints Pruning Results . . . . .	180
8.8	Structural Constraints Pruning Results . . . . .	181
8.9	Initial Meta-Group for Automatic Design . . . . .	193
8.10	Initial Meta-Molecules for Automatic Design . . . . .	196
8.11	Meta-Molecules Surviving Structural Pruning . . . . .	199
8.12	Final Display of Meta-Groups Tree after Automatic Design . . . . .	228
9.1	Interactive Design Section Preparation Configuration Screen . . . . .	231
9.2	Interactive Design Section Design Configuration Screen . . . . .	232
9.3	Interactive Refrigerant Design Preparation . . . . .	235
9.4	Interactive Constraint Allocation . . . . .	236
9.5	Interactive Design Space Solution Grid . . . . .	237

9.6	Legend for Design Space . . . . .	238
9.7	$-\text{CH}_3$ Temporary Group Vector . . . . .	240
9.8	Molecule Within the Target Region . . . . .	241
10.1	Evaluation Section Specifications Configuration Screen . . . . .	261
10.2	Evaluation Section Specifications Configuration Screen . . . . .	263
10.3	Display of Initial $P_{vp}$ Estimation Procedure . . . . .	265
10.4	Display of Intermediate $P_{vp}$ Estimation Procedure . . . . .	267
10.5	Display of Final $P_{vp}$ Estimation Procedure . . . . .	268
10.6	Estimated $P_{vp}$ for Dichlorodifluoromethane . . . . .	270
10.7	Values Configuration with Estimated Physical Properties . . . . .	279
10.8	Estimated $P_{vp}$ vs. $T$ Plot . . . . .	280
10.9	Estimation Technique Choices after $P_{vp}$ Technique Selection . . . . .	284
10.10	Estimating $P_{vp}$ Values for Three Molecules . . . . .	291

# List of Tables

1.1	Sections of the Molecular Design System	4
3.1	System Sections and Configurations	18
4.1	Property Object Instance Variables: Critical Temperature Property Instance	28
4.2	All Properties	30
4.3	All Property Classes	31
4.4	Constraint Object Instance Variables	31
5.1	Data-Bank Physical Properties	48
5.2	Data Base Molecule Instance Variables	55
5.3	Data Object Instance Variables	56
5.4	Data Base References	57
5.5	Data Base Keyword Instance Variables	58
5.6	Keyword Function Instance Variables: $T_{br}$ Keyword Function	58
6.1	Molecule Representation Objects	99
6.2	All Atoms Currently Known to the System	99
6.3	Molar Magnetic Susceptibility Contributions	131
7.1	All Estimation Techniques	139
8.1	20 Automatically Designed Molecules	226
8.2	Seven Surviving Automatically Designed Molecules	227
A.1	Complex Groups' Linear Names	300

A.2 Bond Symbols . . . . .	300
----------------------------	-----

# Notation

## Physical Properties

$C_{pL}$	Liquid Heat Capacity
$C_{pL}$	Liquid Heat Capacity
$C_{pS}$	Solid Heat Capacity
$C_{pV}$	Ideal Gas Heat Capacity
$C_{pV}^o$	Ideal Gas Heat Capacity
$C_{pV,298}^o$	Ideal Gas Heat Capacity at 298K
$C_{pV,a}$	Constant Coefficient in Ideal Gas Heat Capacity Cubic Fit
$C_{pV,b}$	Linear Coefficient in Ideal Gas Heat Capacity Cubic Fit
$C_{pV,c}$	Quadratic Coefficient in Ideal Gas Heat Capacity Cubic Fit
$C_{pV,d}$	Cubic Coefficient in Ideal Gas Heat Capacity Cubic Fit
$F_1$	Factor 1
$F_2$	Factor 2
$F_3$	Factor 3
$\Delta G_{f,298}^o$	Standard Gibbs Energy of Formation at 298K
$\Delta H_{f,298}^o$	Standard Enthalpy of Formation at 298K
$\Delta H_m$	Enthalpy of Fusion

$\Delta H_v$	Enthalpy of Vaporization
$H_v$	Enthalpy of Vaporization
$\Delta H_{vb}$	Enthalpy of Vaporization at $T_b$ .
$H_{vb}$	Enthalpy of Vaporization at $T_b$ .
$M_w$	Molecular Weight
$n_A$	Number of atoms
$\omega$	Acentric Factor
$P_c$	Critical Pressure
$P_c^*$	Modified Critical Pressure
$P_{LL}$	Molar Polarizability
$P_{vp}$	Vapor Pressure
$P_{vp}$	Vapor Pressure
$T_b$	Normal Boiling Point
$T_{br}$	Reduced Boiling Point: $T_b/T_c$
$T_{br}^*$	Modified Reduced Boiling Point
$T_c$	Critical Temperature
$T_g$	Glass Transition Temperature
$T_m$	Normal Melting Point
$V_c$	Critical Volume
$V_s$	Solid Volume
$V_v^{sat}$	Saturated Vapor Volume
$Z_c$	Critical Compressibility

**Foreign Symbols**

$\delta_p$  ..... Dipole Moment

$\eta_L$  ..... Liquid Viscosity

$\omega$  ..... Acentric Factor

**Computer Symbols**

**c** ..... control modifier key

**h** ..... hyper modifier key

**m** ..... meta modifier key

**s** ..... super modifier key

**sh** ..... shift modifier key

**sy** ..... symbol modifier key

# Chapter 1

## Introduction

My molecular design system consists of approximately 20,000 lines of LISP code with an additional 17,000 line databank. I implemented the system in Symbolics Common LISP on a Symbolics LISP Machine. In this volume I describe the implementation in sufficient detail to enable use, maintenance, and extension of the system.

### 1.1 Symbolics Environment

My implementation is based on Symbolics' Genera 7.1 development environment. Understanding Genera is needed for maintaining and extending the system. Appendix B provides an outline for becoming familiar with LISP and the Symbolics environment.

### 1.2 System Organization

The system is divided into eight sections corresponding closely to the parts of the methodology described in Volume 1. Each section is composed of one or two subsections

called configurations. Sections are conceptual entities associated with a stage of the methodology. The configuration (actually referring to a configuration of windows) is the actual display you see on the screen. Figure 1.1 shows an example configuration.

The system divides the screen into several *window panes*. These panes are arranged in a *tile* format. This means that individual window panes are not meant to be rearranged or modified in size or shape. The name of a window pane is displayed in a black rectangle at the top of the pane. Not all panes have names.

Three important types of panes occur in every configuration:

**Title Pane:** The Title Pane is displayed along the top of the configuration. It displays the name of the current configuration. The Title Pane in Figure 1.1 displays the name:

**Data Base Section: Plot Configuration.**

**Commands Menu:** The Commands Menu displays commands applicable to the current configuration. Commands are displayed in alphabetical order. Each configuration has its own commands menu. Figure 1.1 shows the Plot Configuration's Commands Menu in the lower right corner of the screen. Command Menus' names are not displayed.

**Interaction Pane:** There is one Interaction Pane for the system. All configurations share the same pane. The Interaction Pane accepts input and displays output. The system prompts for input by displaying the **MD Command:** prompt.

The eight sections of the system are shown in Table 1.1 with their associated configurations. Each section is designed to focus on the accomplishment of a few specific

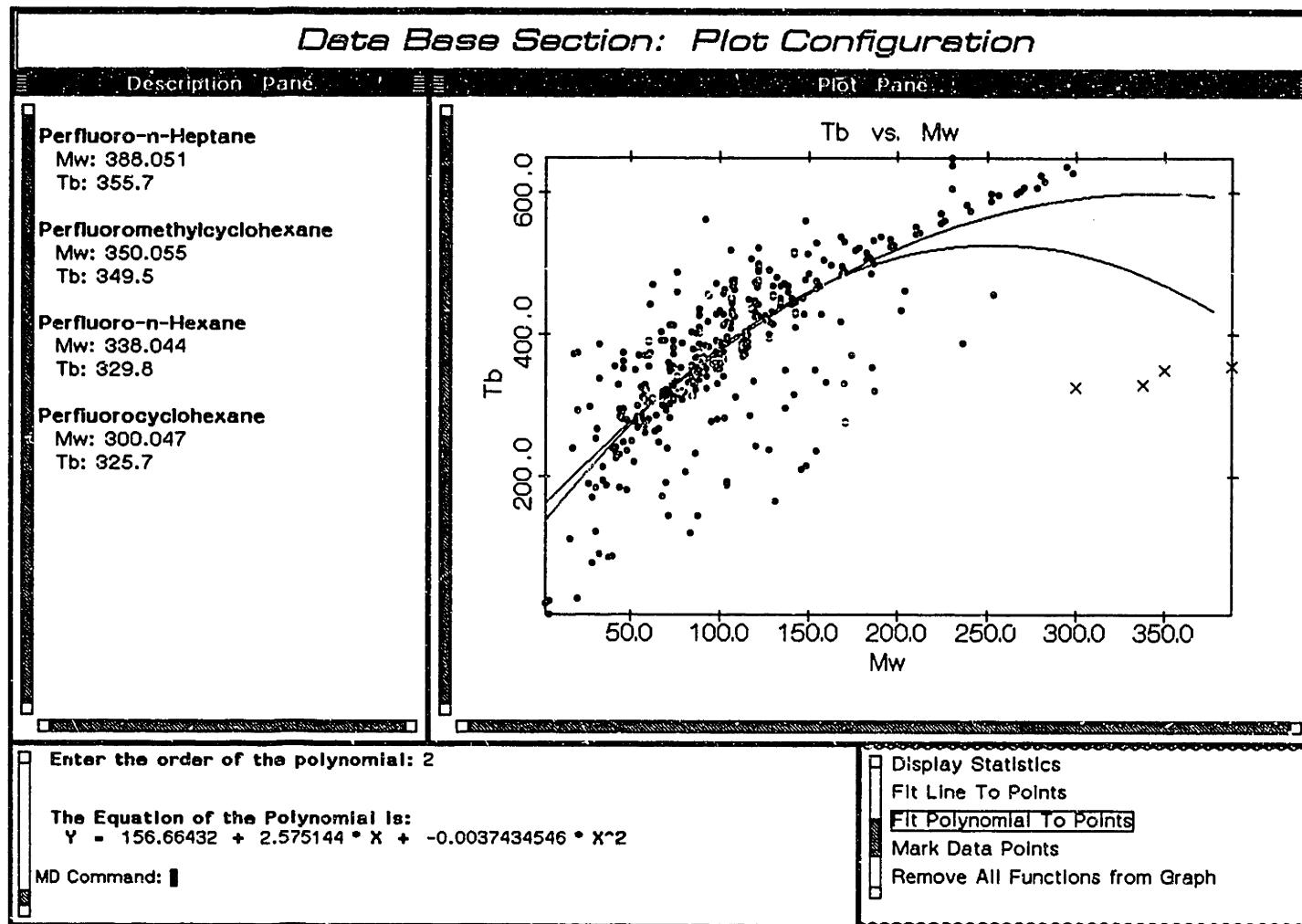


Figure 1.1: Example Configuration: Plot Configuration

Table 1.1: Sections of the Molecular Design System

---

**Login Section**

    Login Section Configuration

**Problem Formulation Section**

    Problem Formulation Section Constraints Configuration

**Data Base Section**

    Data Base Section Plot Configuration

    Data Base Section Data Configuration

**Target Transformation Section**

    Target Transformation Section Configuration

**Interactive Design Section**

    Interactive Design Section Design Configuration

    Interactive Design Section Preparation Configuration

**Automatic Design Section**

    Automatic Design Section Meta Molecules Configuration

    Automatic Design Section Meta Groups Configuration

**Molecule Evaluation Section**

    Molecule Evaluation Section Specification Configuration

    Molecule Evaluation Section Configuration

**Group Contribution Section**

    Group Contribution Section Model Entry Configuration

    Group Contribution Section Editing Configuration

---

tasks. The major tasks throughout all sections are summarized into three classes:

1. Creation/Deletion of objects.
2. Storage of information into objects or retrieval of information from objects.
3. Coercion or transformation of an object of one type into an object of another type.

These three tasks involve the two classes of code which compose most of the system: objects and commands.

## 1.3 Objects

Objects have three important characteristics:

1. Slots or instance variables for information storage.
2. Presentations which specify how an object is displayed and accepted by the user.
3. Methods which are functions specific to each object. The interface to these methods is primarily through commands.

The graph, data points, and functions displayed in Figure 1.1 are all represented in the computer as objects. Data points have instance variables which store their x and y coordinates and associated molecule. A data point is presented either as a filled circle, •, or a cross, ×. Data points have methods for drawing and describing themselves. The graph and function objects have similar methods. However, whereas the description method for a data point returns its x-coordinate, y-coordinate, and associated molecule, the description method for a graph returns its x-axis, y-axis, and number of data points. Applying the same method to different objects can yield very different results.

## 1.4 Commands

Commands manipulate objects and the information they contain. The **Fit Polynomial To Points** command highlighted in Figure 1.1 prompts the designer for a **2-d-active-graph** and the order of the polynomial. The command then extracts all the unmarked data points of the graph and regresses a polynomial of the given order. This polynomial is then presented in the graph as a **function-object**.

Commands provide the interface between the designer and the LISP code implementing the system. The designer activates commands by mousing on their displayed name or typing their name into the interaction pane.

## 1.5 Structure

Chapters describing sections of the implementation are divided into six parts:

1. **Section Layout:** Describes the layout of windows composing the section. The purpose of each window is described.
2. **Section Operation:** Briefly describes the tasks accomplished by the section. This is done at an abstract level. A detailed example is given in Example Usage.
3. **Section Objects:** Describes the major objects used in the section. Important instance variables and methods for some of these objects are described.
4. **Section Commands:** Describes each of the commands relevant to the section.
5. **Section Discussion:** Discusses the concepts used in the section's implementation. I typically present recommendations for improvement.
6. **Example Usage:** Presents a detailed example of the section's use. The goal is to provide step-by-step instruction requiring no prior experience with the system. Designing refrigerants is used to demonstrate many of the sections.

# Chapter 2

## System Basics

This chapter overviews basic operations of the Symbolics environment. I introduce some of the terminology and commands used in following chapters. The description given here is sufficient to run the system. Understanding LISP and the Symbolics environment is necessary to maintain and extend the system. Appendix B provides an outline for learning LISP and the Symbolics environment.

### 2.1 Keyboard

The 88 keys on the Symbolics keyboard are classified into three groups: 1) character keys; 2) function keys; 3) modifier keys. Function keys, e.g., TAB, RUBOUT, and character keys, e.g., a, b, c, are designated by white labels. When a function key is depressed some action occurs, e.g., a tab is inserted or the character to the left of the cursor is deleted. When a character key is depressed a character is inserted at the current location.

Modifier keys are intended to be held down while a function or character key is typed. Modifier keys “modify” the behavior of function and character keys. The most commonly used modifier key is the shift key which causes a capital letter to be typed when held down in conjunction with a character key. Modifier keys are denoted by black labels. Six modifier keys are commonly used. These have the following abbreviations:

**c** – control

**s** – super

**h** – hyper

**sh** – shift

**m** – meta

**sy** – symbol

Thus **c-m-a** means while holding down both the **control** and **meta** modifier keys depress the character key **a**.

Modifier keys are often used in conjunction with mouse gestures. Mouse gestures are discussed next.

## 2.2 Mouse

Symbolics uses a three button mouse. The mouse buttons are designated as **left**, **middle**, and **right**. Clicking a mouse button generates a “mouse gesture”. The verb “mousing” is often used to denote the clicking action. The system developer associates LISP functions with a particular mouse gesture. Thus instead of calling a particular system function to display the system menu, you can mouse right twice.

Symbolics’ presentation system makes objects displayed on the screen mouse sensitive. Mouse sensitivity means that the system recognizes the object or objects currently under the mouse cursor. A typical term used throughout my thesis is to mouse “on”

an object. This means to move the mouse cursor over the displayed object and click one of the mouse buttons.

Modifier keys also effect mouse gestures. Mouse **h-sh-left** means to hold down both the **hyper** and **shift** keys while clicking the left mouse button. The mouse documentation line displays the currently applicable mouse gestures. The documentation line is located at the bottom of the screen just above the status line.

## 2.3 Non-Displayed Commands

A number of commands in my molecular design system are available only through mouse gestures or keyboard entry. This provides the convenience of access but necessitates describing the commands to ensure the user is aware of them.

The action a mouse gesture takes depends on the type of object the mouse is over. Thus for each of the commands listed below, I specify the object types for which the gesture is applicable. For example, to change to another configuration of the system you could type the command

### Select Section

or place the mouse so it is not over any object and click the right mouse button.

**Select Section:** This command is activated by moving the mouse cursor to an empty area of the screen and mousing **right**. A menu is exposed listing all the configurations of the molecular design system. Choosing a configuration exposes it.

**Previous Section:** This command is activated only through keyboard entry. The previous configuration is exposed.

**Input Values:** This command is activated by moving the mouse cursor over any molecular design system object and mousing **h-sh-right**. A menu is exposed displaying the values for some of the object's instance variables. The instance variables displayed are object dependent. These values can be modified.

This facility is rarely used and I suggest removing it. The input of values for instance variables requires a more developed interface than a simple menu.

**Redisplay Pane:** This command is activated by moving the mouse cursor to an empty area of a window pane and mousing **h-sh-right**. The window pane the mouse cursor is over is redisplayed. Each window pane of my molecular design system has its own redisplay function. The redisplay action is thus window pane dependent.

**Delete Self:** This command is activated by moving the mouse cursor over any molecular design system object and mousing **h-right**. The object is deleted from the window pane. Some objects are not deleteable. Attempting to delete a non-deleteable object does nothing.

**Move Object:** This command is activated by moving the mouse cursor over any **graphic-based-object** and mousing **h-left**. The mouse "picks up" the object. In actuality the object is deleted from its window pane and the mouse cursor is changed to resemble its display. The designer repositions the object by moving the mouse.

Mousing left deposits the object at the repositioned mouse location. This command is primarily used for moving atoms, groups, and graphs.

**Displace from Keyboard:** This command is activated by moving the mouse cursor over any **graphic-based-object** and mousing **h-s-right**. This command is analogous to the **Move Object** command. However, repositioning is done through keyboard commands not with the mouse. This command is used when accurate positioning of objects is required. Accurate positioning with the mouse is difficult. The four keyboard commands used are:

- c-F** moves the graphic object forward to the right
- c-B** moves the graphic object backward to the left
- c-P** moves the graphic object up
- c-N** moves the graphic object down

The unit of movement is 1 pixel. This command is primarily used for repositioning atoms.

The following two commands select and deselect objects. Selection is a programming concept made popular by the Macintosh interface[3]. The input for certain commands is taken from a list of selected objects. Symbolics's paradigm would be to prompt for a sequence of objects. I believe this is the better choice except when the list is long or objects choice is based on some complex criteria. The selected object is graphically altered to denote it has been selected. This is usually done by displaying the object in inverse video. Selection is one way to "mark" one or more objects so functions apply only to them.

**Select/Deselect Object:** This command is activated by moving the mouse cursor over any molecular design system object and mousing **h-s-left**. The object typically is redisplayed in inverse video. If the object is already selected then it is deselected.

**Window Select/Deselect:** This command enables the designer to select more than one object at a time. The mouse cursor is used to form a box around those objects the designer wishes to select. The command is activated by moving the mouse cursor to an empty area of the window pane containing the objects to be selected and mousing **h-s-left**. The mouse cursor changes to an upper left corner. Mousing left places this upper left corner at the current mouse location. The mouse cursor then changes to a lower right corner. A “rubber banding box” connects the two corners. Mousing left places this lower right corner at the current mouse location. All objects whose centers are within the box are selected.

### 2.3.1 Symbolics' Commands

I added the following Symbolics system commands to the molecular design system's command table:

- Clear Output History
- Copy Output History into Editor
- Hardcopy File
- Load File
- Compile File.

# Chapter 3

## Login Section

The Login Section is the first configuration the designer sees when the system begins operating. The Login Section provides documentation for all of the system's sections and configurations. It is intended to be a location the designer can retreat to if he or she gets lost. The Login Section is thus analogous to the Home Card in Apple's HyperCard[5]. However, since any configuration of the system is reachable from any other configuration, the Login Section does not serve as a central exchange. There is no requirement to return to the Login Section "on the way" to another configuration.

The Login Section consists of a single configuration. The main task of this configuration is to provide documentation on the system, sections, and configurations.

### 3.1 Section Layout

The screen layout of the Login Section is shown in Figure 3.1. The screen real estate is used by five panes:

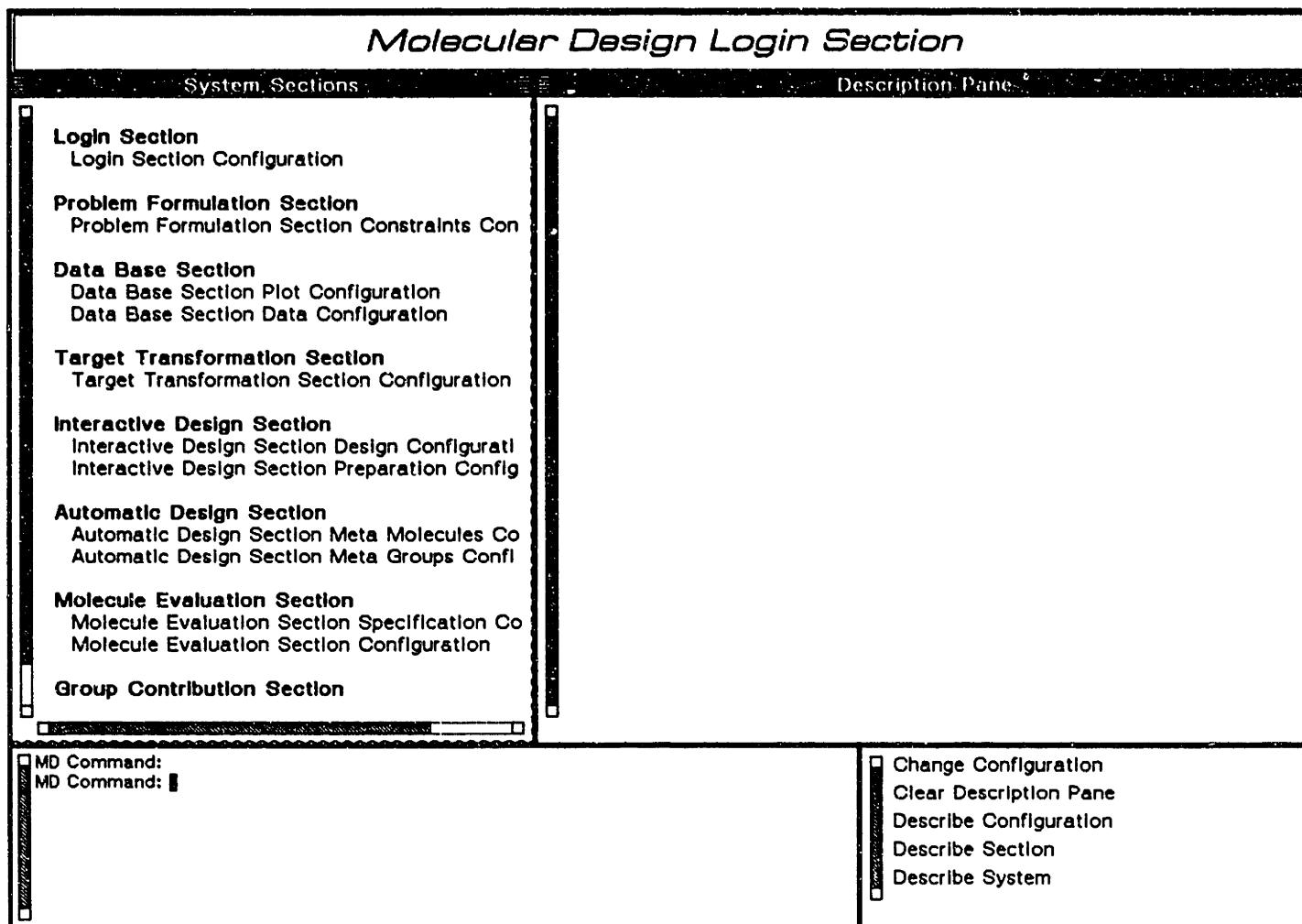


Figure 3.1: Login Section Screen

**Login Section Title Pane:** Displays the title of the Login Section.

**System Sections Pane:** Displays all the sections and configurations of the system.

The display is arranged by section. Each section is composed of one or two configurations. Sections are displayed in bold with their one or two configurations displayed beneath them.

**Description Pane:** This pane displays the requested documentation. New documentation is appended to the end of the pane.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Login Section Commands Menu:** The command menu displaying commands relevant to the Login Section.

## 3.2 Section Operation

The primary task of the Login Section is to provide a central location for system documentation. Documentation can be viewed for a section, configuration, or the system in general. The **Describe System** command displays brief documentation about the molecular design system. The **Describe Section** command prompts for a section. A choice can be made by mousing on one of the sections displayed in the System Sections Pane or typing the section's name. The chosen section's documentation is displayed in the Description Pane. The **Describe Configuration** command operates

analogously. Figure 3.2 shows the Login Section after the Problem Formulation Section was documented.

New documentation is appended to the bottom of the Description Pane. When the displayed documentation becomes excessive the **Clear Description Pane** command clears all displayed documentation.

The **Change Configuration** command prompts the designer for a new configuration. A choice can be made by mousing on one of the configurations displayed in the System Sections Pane. The system changes to the chosen configuration.

### 3.3 Section Objects

The two major objects used in the Login Section are:

1. **Section**
2. **Configuration**

Table 3.1 lists all the sections and configurations known to the system. These objects and associated functions are defined in the file:

`molecular-design:login-section;objects.lisp.`

The **section** object has two instance variables inherited from the basic flavor: **text-based-object**. These are:

1. **pretty-name**
2. **documentation-object**

The **pretty-name** instance variable contains the name of the section. This is the name displayed in the Systems Sections Pane. The **documentation-object** contains a

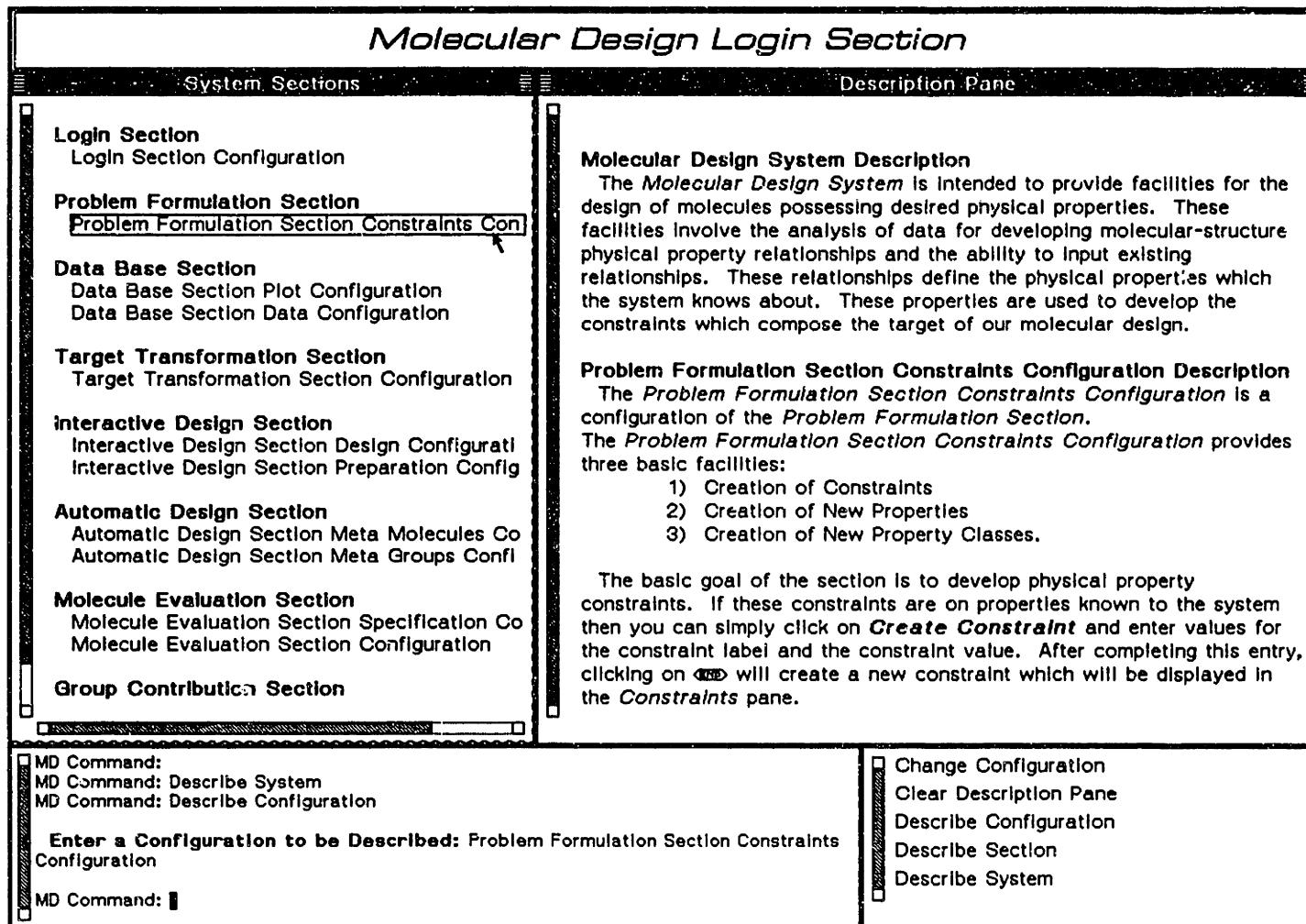


Figure 3.2: Login Section Documentation Facility

Table 3.1: System Sections and Configurations

---

**Login Section**

    Login Section Configuration

**Problem Formulation Section**

    Problem Formulation Section Constraints Configuration

**Data Base Section**

    Data Base Section Plot Configuration

    Data Base Section Data Configuration

**Target Transformation Section**

    Target Transformation Section Configuration

**Interactive Design Section**

    Interactive Design Section Design Configuration

    Interactive Design Section Preparation Configuration

**Automatic Design Section**

    Automatic Design Section Meta Molecules Configuration

    Automatic Design Section Meta Groups Configuration

**Molecule Evaluation Section**

    Molecule Evaluation Section Specification Configuration

    Molecule Evaluation Section Configuration

**Group Contribution Section**

    Group Contribution Section Model Entry Configuration

    Group Contribution Section Editing Configuration

---

documentation string. The **Document Section** command displays this string in the Description Pane.

The **configuration** object has four important instance variables:

1. **Program-Name**
2. **Section**
3. **Pretty-Name**
4. **Documentation-Object**

**Program-name** stores the name of the configuration used by the program. This is typically the **pretty-name** of the configuration with spaces replaced by hyphens. This name is used in the program frame's `:set-configuration` method. The **section** instance variable stores this configuration's associated section. This is used to categorize configurations for display in the System Sections Pane. The **pretty-name** instance variable stores the name of the configuration. The **documentation-object** instance variable stores a documentation string describing the purpose of the configuration.

## 3.4 Section Commands

The following commands are listed in the command menu of the Login Section. The commands' definitions are in the file:

```
molecular-design:login-section;commands.lisp.
```

**Change Configuration:** Prompts the designer for a configuration. This configuration can be entered by typing the name appearing in the System Sections Pane or by mousing on the displayed name. The Login Section Configuration is deexposed and the chosen configuration is exposed.

**Clear Description Pane:** Removes all the displayed documentation from the Description Pane. Clearing the Description Pane is useful when a great deal of documentation has been displayed.

**Describe Configuration:** Prompts the designer for a configuration to be described. The configuration's name can be typed or its displayed name can be moused on. The documentation for the chosen configuration is displayed in the Description Pane.

**Describe Section:** Prompts the designer for a section to be described. The section's name can be entered or its displayed name can be moused on. The documentation for the chosen section is displayed in the Description Pane.

**Describe System:** Displays documentation about the overall system in the Description Pane.

## 3.5 Section Discussion

The documentation facility provided by the Login Section is fully functional. The documentation available for the sections and configurations is meager when present. More documentation must be added.

Examples of the system operating provide greater assistance to the designer than any written documentation could. Session "recordings" could be replayed to demonstrate actual mouse motions, command executions, and object creation and display. Implementation of such a recording facility requires intercepting commands, tracking

mouse motions, and detecting mouse gestures. I did not work on these ideas in any depth but am quite sure all are possible. Symbolics' user interface paradigm provides a good basis for such a recording system.

## 3.6 Example Usage

The system exposes the Login Section when started. The purpose of the Login Section is to provide documentation on the system, sections, and configurations.

**Action 3.1** *Mouse left on the Describe System command.*

Some extremely brief documentation on the Molecular Design System is displayed in the Description Pane.

The System Sections Pane displays the system's sections and configurations. Each is documented.

**Action 3.2** *Mouse left on the Describe Configuration command.*

The system prompts for a configuration:

**Enter a Configuration to be Described:**

**Action 3.3** *Mouse left on the Problem Formulation Section Constraints Configuration.*

The documentation for the configuration is appended to the end of the Description Pane.

# Chapter 4

## Problem Formulation Section

The first step of any design is to establish the target. In molecular design our target consists of constraints on structural, physical, and chemical properties important to the performance of the desired chemical product. The main purpose of the Problem Formulation Section is to provide an interface with which the designer can enter physical property constraints. The Problem Formulation Section consists of a single configuration.

### 4.1 Section Layout

The screen layout of the Problem Formulation Section is shown in Figure 4.1. The screen real estate is used by five panes:

**Property List Pane:** Displays all the physical properties known to the system. The physical properties are arranged by property class.

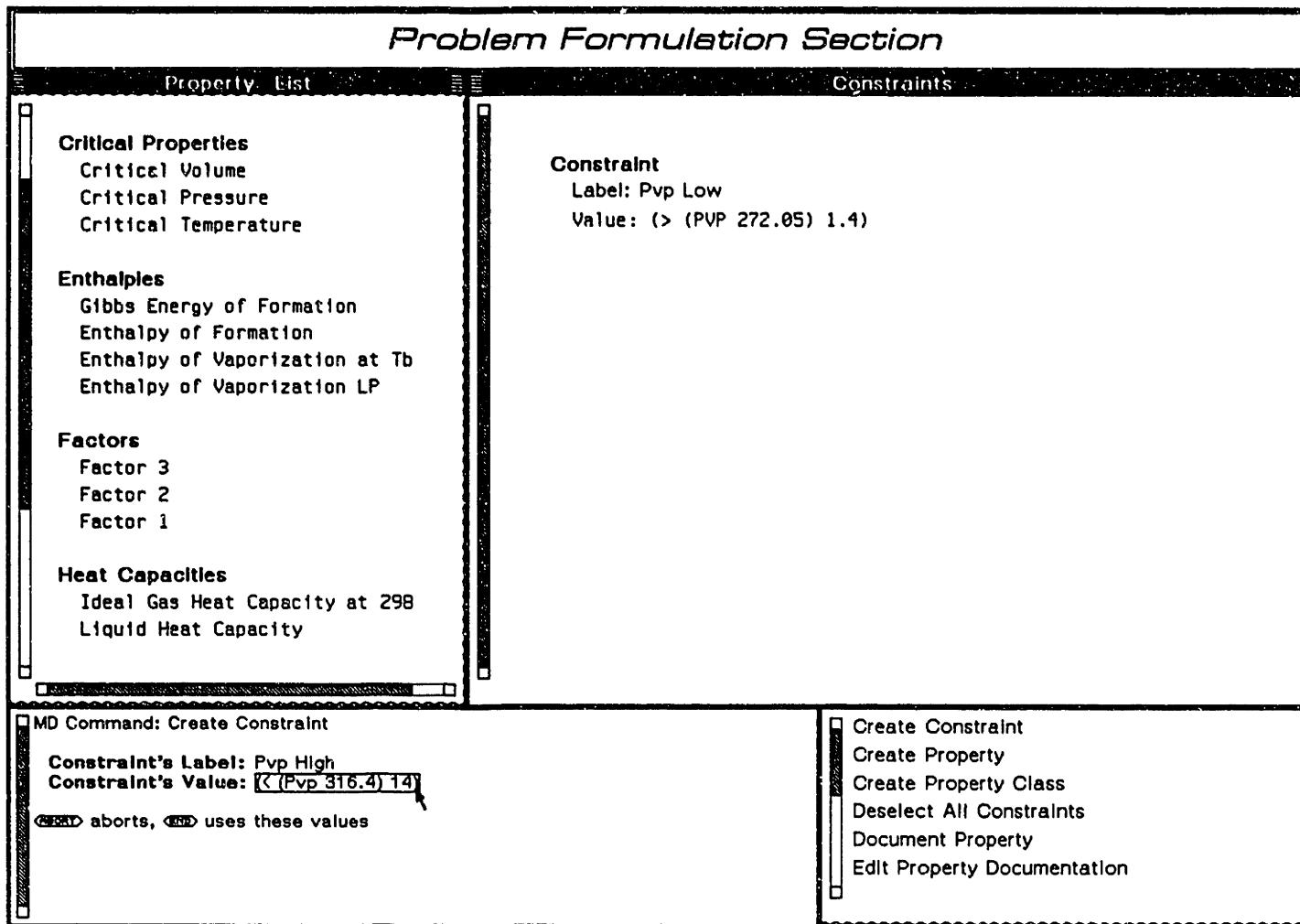


Figure 4.1: Problem Formulation Section Screen

**Constraints Pane:** Displays the constraints created by the designer. Constraints displayed in this window can be edited, selected, and deleted.

**Problem Formulation Section Title Pane:** Displays the title of the Problem Formulation Configuration.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Problem Formulation Section Commands Menu:** The command menu containing the commands relevant to the Problem Formulation Section.

## 4.2 Section Operation

The major task of the Problem Formulation Section is to provide an interface through which the designer enters physical property constraints. Two subtasks of the section enable expansion of the system's knowledge of physical properties and physical property classes.

### 4.2.1 Creating Constraints

Physical property constraints are collections of mathematical operations and physical properties. Constraints evaluate to either true or false when values are specified for the physical properties. The designer enters a constraint using the **Create Constraint** command. Activating this command exposes a menu in the Molecular Design Interac-

tion Pane prompting for the two items of information needed to create a constraint:

1. Constraint Label
2. Constraint Value

The constraint label is a short string which identifies the constraint. A constraint's label often conveys the intent of the constraint. Thus a constraint which is looking for high values for the normal boiling point might be named "Tb Large".

The constraint value is the actual equation which is evaluated to test constraint satisfaction. Any function of physical properties is allowed to be used as a constraint value. The two restrictions are that the system must know all the physical properties used in the value and the value must return either `t` or `nil`. Knowledge of physical properties means that the symbols used in the constraint value must be understood by the system to represent physical property values. The system defines two or three symbols which represent each physical property. These are the symbols which must be used when entering a constraint value. The documentation for any physical property states what these symbols are. The **Document Property** command displays the documentation for a chosen physical property.

After entering the label and value the system creates a constraint object adding it to the Constraints Pane. Constraints displayed in the constraint pane can be edited, selected, and deleted.

The **Verify Constraints** and **Transform Constraints** commands check and prepare the entered constraints for further processing. The designer is free to enter any expression for the value of a constraint. However, the system must be able to parse this expression into a valid LISP form. The **Verify Constraints** command parses each

of the selected constraints checking that each of the symbols used is either a number, function, or physical property symbol. Once all constraints are verified the **Transform Constraints** command coerces each of the property constraints into a transformed constraint object, adds these objects to the Transformed Constraints Pane of the Target Transformation Section, and exposes that section.

#### 4.2.2 Creating Properties

If the entered constraint contains a symbol which is not recognized as a function or a physical property symbol, the constraint is not allowed to be further processed. Thus to use new properties in a constraint the system must first be informed about the new property.

Entering a new property begins with identifying whether or not the property belongs to one of the existing property classes. If an appropriate class does not exist then it must be created first. The creation of property classes is discussed in the next section.

Once an appropriate property class is present the new property can be entered. A property object contains a variety of information about the property. The important instance variables are:

1. **pretty-name**
2. **short-name**
3. **variable-dependencies**
4. **default-minimum**
5. **default-maximum**
6. **property-class**

The **Create Property** command prompts for values for each of these variables, creates a property object, and adds this object to the Property List Pane.

The entry of a new property necessitates entering one or more estimation techniques to estimate the property. The system can currently use either group contribution or equation oriented estimation techniques. The **Create Estimation Technique** command prompts for a physical property and changes to the Group Contribution Section: Model Entry Configuration. The Group Contribution Section provides facilities for the entry of both equation oriented and group contribution techniques. These facilities are detailed in Chapter 6.

#### 4.2.3 Creating Property Classes

When a new class of physical properties are to be entered it is necessary to first create a new property class. Property classes are used to organize the display of physical properties. The display of physical properties in the Property List Pane of Figure 4.1 is organized by property class. The **Create Property Class** command prompts for the class's **label-string**, creates the property-class object and adds the object to the Property List Pane.

### 4.3 Section Objects

The three major objects used in the Problem Formulation Section are:

1. **property**
2. **property-class**
3. **constraint**

Table 4.1: Property Object Instance Variables<sup>†</sup>

Instance Variable	Example Value
Pretty-Name	“Critical Temperature”
Short-Name	“Tc”
Argument-Symbol	Tc
Character	“π”
Variable-Dependencies	nil
Default-Minimum	200
Default-Maximum	1000
Property-Class	“Critical Properties”
Documentation	“The default minimum of ....”

<sup>†</sup> Critical temperature property instance.

I discuss the important instance variables and methods for each of these objects.

### 4.3.1 Property Object

The **property-object** stores information about a physical property. The important instance variables of the property object are shown in Table 4.1 with example values.

I briefly describe the purpose of each of these instance variables.

Properties are used in the expressions composing a constraint’s value. To facilitate entry a short name is acceptable for input. This short name is a string stored in the **short-name** instance variable. To write a constraint that the critical temperature should be greater than 400 K, one uses the short name for the critical temperature, Tc, and writes

$(> \text{Tc} 400)$ .

The **character** instance variable is used exactly like the short name. Please note that the name, “Character”, is deceiving. A string is required, not a character. Orig-

nally I constructed a special font which contained many of the physical property symbols. This was not a simple procedure and I no longer recommend it. Thus instead of constructing a symbol which looks like  $T_c$ , I use `Tc`.

The `variable-dependencies` instance variable stores a list of the state variables upon which this property is dependent. The state variables are *Temperature* or *Pressure* or both.

The `default-minimum` and `default-maximum` instance variables store lower and upper limits for this property. The limits are used for determining the limits on the axes in an interactive design.

Every property belongs to a property class. The `property-class` instance variable stores the name of this class. Property classes are used to present the properties in an organized manner in the Property List Pane.

The property flavor definition and its associated function definitions are in the file:

`molecular-design:problem-formulation-section;objects.lisp`.

Table 4.2 displays all the properties currently known to the system. The definitions for these instances are in the file:

`molecular-design:properties;property-instances.lisp`.

### 4.3.2 Property Class Object

The property class object is used to organize the properties known to the system. Each property instance has a property class instance stored in its `property-class` instance variable. When presented to the user the properties are grouped by property

Table 4.2: All Properties

---

Critical Temperature	Critical Pressure
Critical Volume	Vapor Pressure
Acentric Factor	Enthalpy of Vaporization Low Pressure
Enthalpy of Vaporization at Tb	Enthalpy of Formation
Gibbs Energy of Formation	Normal Boiling Point
Normal Melting Point	Reduced Boiling Point
Liquid Heat Capacity	Ideal Gas Heat Capacity at 298
Factor 1	Factor 2
Factor 3	Glass Transition Temperature
Permachor	Volume Resistivity
Molar Polarization	Molar Volume
Molecular Weight	Thermal Conductivity
Solid Heat Capacity	Rao Function
Polar Solubility Parameter	Hydrogen Solubility Parameter

---

class. Properties which have no value for their **property-class** instance variable are assigned to the **Unspecified Properties** property class.

Property class objects have a single important instance variable: **label-string**. The **label-string** is the text which is displayed when a property class object is presented. The **property-class** flavor definition and its associated function definitions are in the file:

**molecular-design:problem-formulation-section;objects.lisp.**

Table 4.3 shows all the property classes currently known to the system. The definitions for these instances are in the file:

**molecular-design:properties;property-class-instances.lisp.**

---

Table 4.3: All Property Classes

---

Critical Properties Class	Pressures Class
Enthalpies Class	Volumes Class
Temperatures Class	Heat Capacities Class
Factors Class	Auxiliary Class

---

### 4.3.3 Constraint Object

The Constraint Object holds the physical property constraints which will be used in the design procedures. The important instance variables of the constraint object are shown in Table 4.4. The constraint flavor definition and associated functions are in the file:

`molecular-design:problem-formulation-section;objects.lisp.`

The four slots:

1. **Label-Style**
2. **Value-Style**
3. **Label-Indent**
4. **Value-Indent**

---

Table 4.4: Constraint Object Instance Variables

---

Instance Variable	Default
label-string	-
value	-
label-style	Swiss.Roman.Large
value-style	Property.Roman.Large
label-indent	2
value-indent	2

---

store information used to present a constraint instance. The `label-style` and `value-style` instance variables contain the character-styles used to display the text. The `label-indent` and `value-indent` instance variables contain the number of spaces this displayed text will be indented. A presentation of a typical constraint is shown in the Constraints Pane of Figure 4.1.

The `value` instance variable holds the actual LISP code which describes the constraint. The interactive design procedure uses this code to identify the feasible region. In the automatic design procedure this code is evaluated for each meta-molecule to identify those meta-molecules which should be retained or pruned.

The `label-string` instance variable stores a string used to identify the constraint. This is used in labeling the constraint when it appears throughout the system. A `label-string` should denote meaning of the constraint. Thus the constraint shown in the Constraints Pane of Figure 4.1 is a requirement that the vapor pressure at 272.05K must be greater than 1.4 bar. This constraint refers to the desire to have the lowest pressure in a refrigeration cycle be greater than 5 psig and is thus appropriately named *Pvp Low*.

## 4.4 Section Commands

The following commands are listed in the Command Menu of the Problem Formulation Section. The definitions of these commands are in the file:

```
molecular-design:problem-formulation-section;commands.lisp.
```

**Create Constraint:** Prompts for the label and value of a property constraint, creates the constraint object, and adds this object to the Constraints Pane.

**Create Property:** Prompts for the following variable values:

1. Pretty Name
2. Short Name
3. Variable Dependencies
4. Default Minimum
5. Default Maximum
6. Property Class

The system uses the entered values to create a property object. This added to the Property List Pane. The pane is redisplayed to show the newly created property.

**Create Property Class:** Prompts for the name of the new class, creates the property class object, and adds this object to the Property List Pane. The pane is redisplayed to show the newly created property class.

**Deselect All Constraints:** Deselects all the constraints displayed in the Constraints Pane. Constraints are selected for verification and transformation.

**Document Property:** All property objects inherit an instance variable which stores documentation. This command prompts for a property object, extracts the documentation from that object, and displays the documentation on a window resource exposed over the Constraints Pane. One of the most important pieces of documentation is the

symbols known by the system to represent the physical property. For example, the system recognizes **vapor-pressure** as a variable for the vapor pressure physical property. However, to simplify entry the system also recognizes **Pvp**.

**Edit Property Documentation:** Enables the designer to enter or edit a property's documentation. Entering documentation should be done after creating a new property.

Documentation is entered using an editor resource. The command prompts for a property object whose documentation is to be edited. The editor resource is exposed over the Constraints Pane. The current documentation is extracted from the object and inserted into the editor. The default documentation for all objects is the string "No Documentation Available".

Pressing the <END> key at the completion of editing causes the edited string to be extracted from the editor and stored back into the property object. I did not implement code to save this new documentation to file. Permanent changes still must be done through the ZMACS editor.

**Enter Estimation Technique:** After creating a new property it is necessary to enter one or more estimation techniques for it. The Group Contribution Section provides facilities for entering equation oriented or group contribution techniques. This command prompts for a property and then changes to the Model Entry Configuration of the Group Contribution Section.

**Redisplay Panes:** Causes the Property List Pane, Constraint Pane, and the Problem Formulation Section Commands Menu to be redisplayed.

**Remove All Constraints:** Deletes all the property constraints displayed in the Constraints Pane. Once constraints are deleted there is no way to retrieve them.

**Restore Properties:** Most objects can be deleted by mousing **h-right** on the object. Sometimes to reduce the number of possibilities to consider, it may be helpful to delete some of the properties displayed in the Property List Pane. Deletion of a property removes it only from the pane not from the system. This command adds all the properties known to the system back into the Property List Pane.

**Save Property to File:** Once a new property has been created it must be saved into a file if it is to be kept for future use. This command prompts for the property object to be saved and then writes a macro to the file:

```
molecular-design:properties;property-instances
```

The macro expands into several definitions when loaded defining the necessary information for recreating the property object.

**Select All Constraints:** Selects all constraints displayed in the Constraints Pane. Constraints are selected for verification and transformation.

**Transform Constraints:** This command collects all the selected constraints from the Constraints Pane, coerces each constraint object into a transformed constraint object, adds these transformed constraints to the Transformed Constraints Pane of the Target Transformation Section, and changes configurations to the Target Transformation Section. This command is one of a number of commands which uses the idea of

object coercion.

**Verify Constraints:** Checks each of the selected constraints to ensure the elements of its value are either numbers, LISP functions, or properties. If a constraint is not verified, it is deselected and its label is presented in a message to the designer.

## 4.5 Section Discussion

The Problem Formulation Section provides a very simple interface for entering physical property constraints. Representing problem specific knowledge and physical property symbols are issues which should be considered in future work.

### 4.5.1 Knowledge Representation

Specifying constraints in a molecular design is a difficult task. A great deal of problem specific knowledge is required to identify the important physical properties. Knowledge of existing compound's physical properties is needed to form reasonable constraints.

The Data Base Section provides assistance in identifying the physical property values of existing compounds. I did not find a representation for problem specific knowledge.

### 4.5.2 Property Symbols

The display of physical property symbols is a difficult problem. Many physical properties such as:

$$\Delta H_{f,298}^{\circ} \quad C_{p_v}^{\circ} \quad V_v^{sat}$$

are very difficult to enter through the keyboard. I created a special font in which each of the characters corresponded to a physical property symbol. This character is stored in the `character` instance variable of the property object.

Creating special characters is a cumbersome procedure. In addition the characters created are only for screen display. Hardcopying a file with these characters results in the system substituting a different font.

I do not have any recommendations on how to solve this problem. Some simple word processors still in use today have the same difficulty. The character style facilities provided in Genera may be able to address some of the problem. Creating larger and smaller characters is possible. However, to make entry easy it would be necessary to have facilities for entering superscripts and subscripts. Such facilities are not yet available.

## 4.6 Example Usage

These instructions detail a step-by-step example usage of the Problem Formulation Section. We accomplish the following tasks:

1. Entering constraints for designing refrigerants.
2. Creating a new property class.
3. Entering a new physical property.

### Entering Constraints

The Problem Formulation Section provides the interface for entering constraints used in both interactive and automatic designs. Constraints are entered as LISP functions

of physical properties. All the physical properties used in a constraint must be known by the system. The only restriction on the constraint function is that it must return `t` or `nil` when physical property values are inserted.

The **Document Property** command provides a list of symbols which are used to represent a particular physical property. For example, our first constraint is a constraint on vapor pressure.

**Action 4.1** *Mouse left on the Document Property command.*

The system prompts for a physical property:

**Enter a physical property:**

**Action 4.2** *Mouse left on Vapor Pressure displayed in the Property List Pane.*

Documentation for vapor pressure is displayed in a window resource exposed over the constraints pane. Part of the documentation states that the variables which the system recognizes as vapor pressure are:

1. `vapor-pressure`
2. `Pvp`

The second variable, `Pvp`, is preferred. Type any character to remove the exposed window resource.

**Action 4.3** *Press the space bar once.*

The variable names for all the physical properties known to the system are found in a similar manner.

We enter the four constraints used in the refrigerant design case study discussed in Volume 1.

**Action 4.4** *Mouse left on the Create Constraint command.*

The system displays an accepting-values menu:

**Constraint's Label:** *some value*

**Constraint's Value:** *some value*

**ABORT** aborts, **END** uses these values

The actual values displayed after the prompts are not important when the menu is initially displayed.

**Action 4.5** *Mouse left on the phrase displayed after the Constraint's Label: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.6** *Type in the label of our first constraint: Pvp Low. Press the return key when you complete the entry.*

**Action 4.7** *Mouse left on the phrase displayed after the Constraint's Value: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.8** *Type in the value of our first constraint: (> (Pvp 272.05) 1.4). Press the return key when you complete the entry.*

When both entries are made the constraint is created and added to the Constraints Pane.

**Action 4.9** *Press the end key.*

We now enter the second constraint.

**Action 4.10** *Mouse left on the Create Constraint command.*

The system displays an accepting-values menu:

**Constraint's Label:** *a constraint's label*

**Constraint's Value:** *a constraint's value*

**ABORT** aborts, **END** uses these values

The actual values displayed after the prompts are not important when the menu is initially displayed.

**Action 4.11** *Mouse left on the phrase displayed after the Constraint's Label: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.12** *Type in the label of our second constraint: Pvp High. Press the return key when you complete the entry.*

**Action 4.13** *Mouse left on the phrase displayed after the Constraint's Value: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.14** *Type in the value of our second constraint: (< (Pvp 316.4) 14). Press the return key when you complete the entry.*

When both entries are made the constraint is created and added to the Constraints Pane.

**Action 4.15** *Press the end key.*

We now enter our third constraint.

**Action 4.16** *Mouse left on the Create Constraint command.*

The system displays an accepting-values menu:

**Constraint's Label:** *a constraint's label*

**Constraint's Value:** *a constraint's value*

**ABORT** aborts, **END** uses these values

The actual values displayed after the prompts are not important when the menu is initially displayed.

**Action 4.17** *Mouse left on the phrase displayed after the Constraint's Label: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.18** *Type in the label of our third constraint: Hv Large. Press the return key when you complete the entry.*

**Action 4.19** *Mouse left on the phrase displayed after the Constraint's Value: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.20** *Type in the value of our third constraint: (> (Hv 272.05) 18.4). Press the return key when you complete the entry.*

When both entries are made the constraint is created and added to the Constraints Pane.

**Action 4.21** *Press the end key.*

We now enter the value for our fourth and final constraint.

**Action 4.22** *Mouse left on the Create Constraint command.*

The system displays an accepting-values menu:

**Constraint's Label:** *a constraint's label*

**Constraint's Value:** *a constraint's value*

**ABORT** aborts, **END** uses these values

The actual values displayed after the prompts are not important when the menu is initially displayed.

**Action 4.23** *Mouse left on the phrase displayed after the Constraint's Label: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.24** *Type in the label of our fourth constraint: Cpl Small. Press the return key when you complete the entry.*

**Action 4.25** *Mouse left on the phrase displayed after the Constraint's Value: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.26** *Type in the value of our final constraint: (< (Cpl 294.2) 32.2). Press the return key when you complete the entry.*

When both entries are made the constraint is created and added to the Constraints Pane.

**Action 4.27** *Press the end key.*

We now have four physical property constraints displayed in the Constraints Pane.

## Creating Property Classes

The physical properties displayed in the Property List Pane are arranged by property class. All physical properties known to the system are assigned to a property class. If a new property is of a different class then it is first necessary to create this property class.

**Action 4.28** *Mouse left on the Create Property Class command.*

The system prompts for the name of the new class:

**Enter the name of the new class:**

At this prompt you should type in the new property class's name.

**Action 4.29** *Type in the name: Environmental Properties. Press the return key when done.*

The system creates a new property-class object and adds it to the Property List Pane.

The system next prompts as to whether or not the new property class should be saved to a file.

**Do you wish to save this class to file?**

You should note that the system provides no facilities for removing property classes once saved. Removing a property class necessitates editing the file:

**molecular-design:properties;property-class-instances.lisp.**

The structure of the file is very simple and easily edited.

**Action 4.30** *Enter No in response to the prompt. Press the return key once the entry is complete.*

## Creating New Properties

Entering new physical properties typically involves creating a new property object in the system and then entering one or more estimation techniques for this new property. The Problem Formulation Section provides a facility for creating the new property object and saving it in a file.

**Action 4.31** *Mouse left on the Create Property command.*

The system exposes a menu prompting for information needed to construct a new property:

**Pretty Name:** *some value*

**Short Name:** *some value*

**Variable Dependencies:** Temperature Pressure

**Default Minimum:** *some value*

**Default Maximum:** *some value*

**Property Class:** *some value*

**ABORT** aborts, **END** uses these values.

The initial values displayed after the prompts are not important.

As an example we enter the values for the polar solubility parameter physical property.

**Action 4.32** *Mouse left on the phrase displayed after the Pretty Name: prompt.*

*The phrase is replaced by a blinking cursor.*

**Action 4.33** *Type in the pretty name of our new physical property: Polar Solubility Parameter.*

**Action 4.34** *Mouse left on the phrase displayed after the Short Name: prompt.*

*The phrase is replaced by a blinking cursor.*

**Action 4.35** *Type in the short name of our new physical property:  $\delta P$ .*

The  $\delta$  symbol is entered by typing `sy-d`. The polar solubility parameter is considered independent of temperature and pressure. No variable dependencies are specified.

**Action 4.36** *Mouse left on the phrase displayed after the Default Minimum: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.37** *Type in the default minimum value of our new physical property: 0.*

**Action 4.38** *Mouse left on the phrase displayed after the Default Maximum: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.39** *Type in the default maximum value of our new physical property: 50.*

**Action 4.40** *Mouse left on the phrase displayed after the Property Class: prompt. The phrase is replaced by a blinking cursor.*

**Action 4.41** *Mouse left on the Environmental Properties class displayed in the Property List Pane.*

When all entries are complete the system creates the new property object and adds it to the property list pane.

**Action 4.42** *Press the <END> key.*

## Transforming Constraints

Once physical property constraints are entered into the Constraints Pane they are ready for further processing. The next steps in their processing are to verify their constraint values and coerce them into **transformed-constraint** objects. The required commands operate only on the selected constraints displayed in the Constraints Pane.

**Action 4.43** *Mouse left on the Select all Constraints command.*

**Action 4.44** *Mouse left on the Verify Constraints command.*

The system checks each constraint's value to ensure all symbols are known. The system displays a message that all constraints were verified.

**All constraints were verified.**

**Action 4.45** *Mouse left on the Transform Constraints command.*

The system collects the selected constraints of the Constraints Pane, coerces each into a **transformed-constraint** object, adds them to the Transformed Constraints Pane of the Target Transformation Section, and then changes configurations to the Target Transformation Section.

# Chapter 5

## Data Base Section

The Data Base Section provides facilities for manipulating and displaying physical property data. The need to examine physical property data for existing compounds occurs in several of the system's sections. In the Problem Formulation Section constraints are often formulated in an evolutionary manner. For example, designing a refrigerant with better property values than Freon-12 requires knowledge of the properties of Freon-12.

The Molecule Evaluation Section can evaluate the accuracy of entered estimation techniques. Results from estimations can be compared to values obtained from the data bank.

The heart of the section is a data-bank containing physical property data on 485 compounds. Properties stored in the data bank are shown in Table 5.1. Units, reference, and temperature dependence are specified for each physical property data value.

The section is divided into two configurations: 1) Data Configuration; 2) Plot Configuration. This division is along the lines of numerical display of data and graphical

Table 5.1: Data-Bank Physical Properties

$T_c$	$P_c$	$V_c$
$Z_c$	$\omega$	$n_A$
$\delta_p$	$C_{p,V}^\circ$	$\eta_L$
$\Delta H_m$	$\Delta G_{f,298}^\circ$	$\Delta H_{f,298}^\circ$
$\Delta H_{vb}$	$T_m$	$T_b$

display of data. The Data Configuration displays data in a tabular format. The Plot Configuration displays data in a graphical format.

## 5.1 Data Configuration Layout

The screen layout of the Data Configuration is shown in Figure 5.1. The screen real estate is used by six panes:

**Data Base Section Data Configuration Title Pane:** Displays the title of the Data Configuration.

**Compounds Pane:** Displays all data base molecules known to the system. There are currently 485 molecules.

**Keywords Pane:** Displays data-base-keywords and keyword functions. Selection of these keywords specifies which properties are accessed from the selected data base molecules.

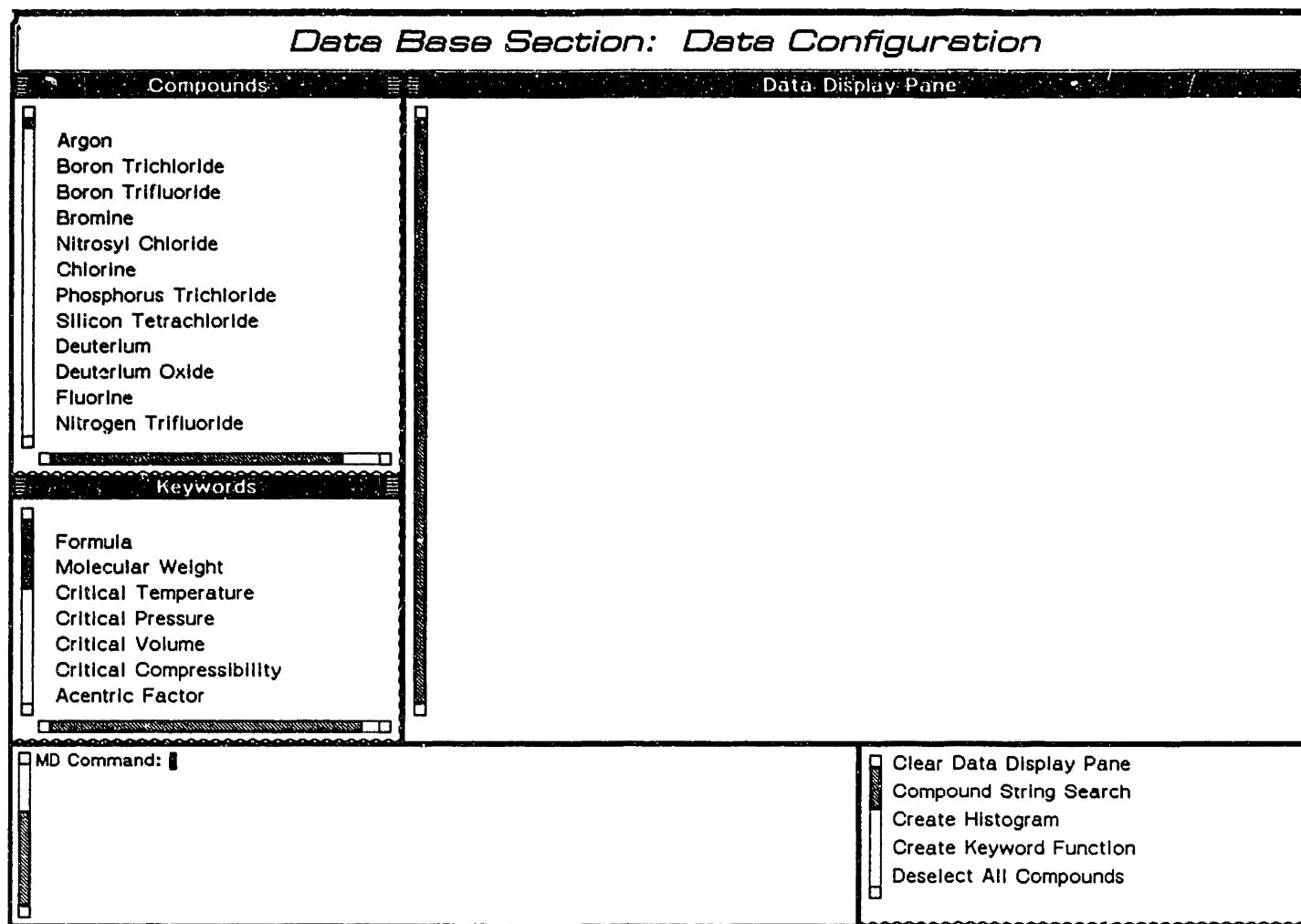


Figure 5.1: Data Base Section Data Configuration Screen

**Data Display Pane:** Displays physical property values for any of the stored compounds in a tabular format. Two tabular forms are used: data display objects and correlation matrices.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and inputs.

**Data Base Section Data Configuration Commands Menu:** The command menu containing commands relevant to the Data Configuration.

## 5.2 Plot Configuration Layout

The physical layout of the Plot Configuration is shown in Figure 5.2. The screen real estate is used by six panes:

**Data-Base Section Plot Configuration Title Pane:** Displays the title of the Plot Configuration.

**Description Pane:** Data points displayed in active graphs and histogram bars displayed in histograms are all “active”. This means that they are actual objects and can be described. The description of a data point or histogram bar is displayed in this pane. A typical use of the description facility is to describe outliers in an x-y plot.

**Plot Pane:** The plot pane displays 2D active graphs and histograms. Objects displayed in this pane are graphic objects capable of being moved and resized.

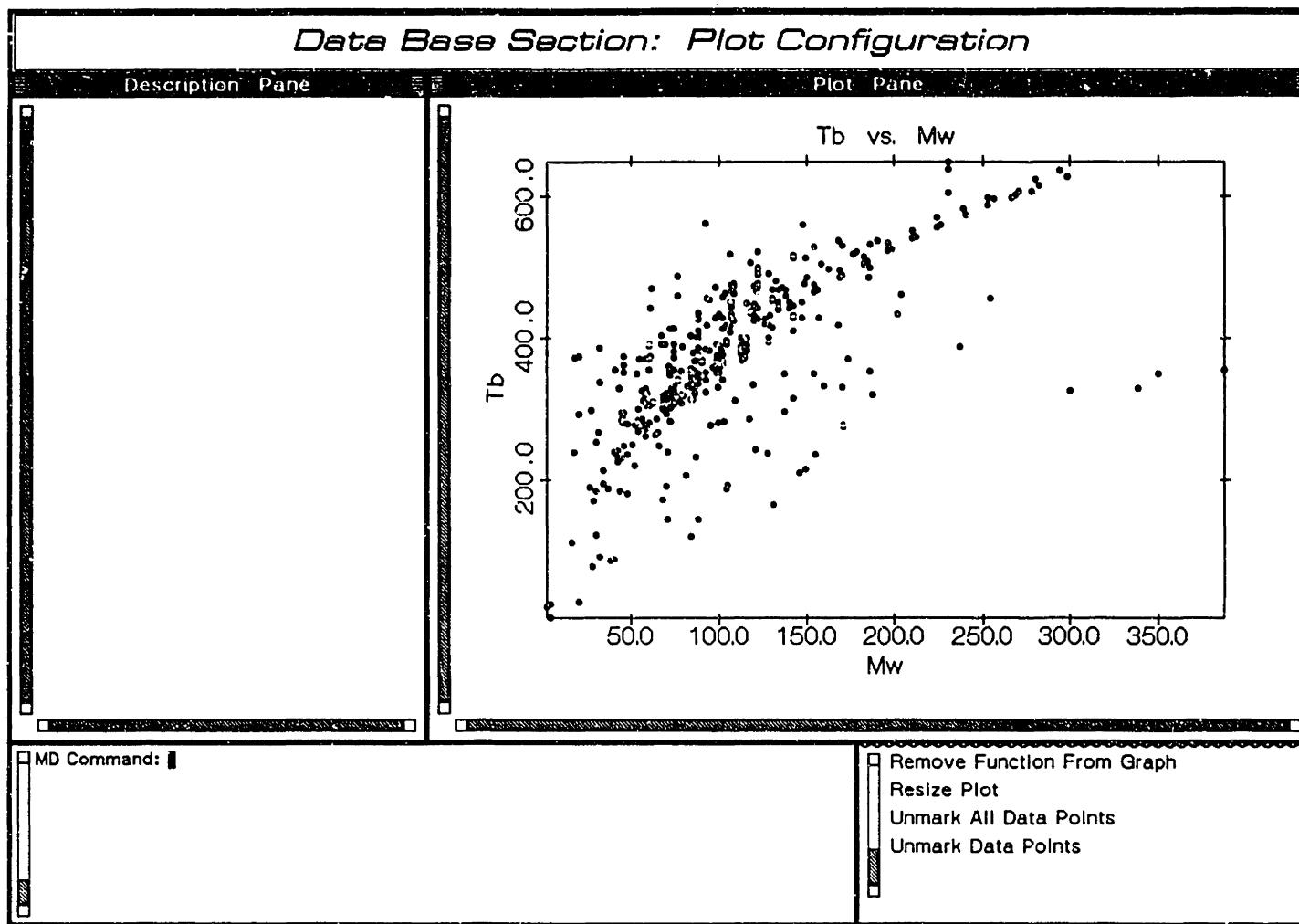


Figure 5.2: Data Base Section Plot Configuration Screen

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Data-Base Section Plot Configuration Commands Menu:** The command menu containing commands relevant to the Plot Configuration.

## 5.3 Section Operation

The examination of physical property data typically begins in the Data Configuration. Physical property data is displayed for a number of molecules. This data can then be displayed graphically in the Plot Configuration.

To display physical property data we need to select some molecules and some keywords. Molecules are displayed in the Compounds Pane. Keywords are displayed in the Keywords Pane. Selecting molecules for display is done in four ways:

1. Selecting all the compounds. This is done using the **Select All Compounds** command.
2. Selecting individual compounds. This is done by the standard selection process. Mousing **h-s-left** on a compound toggles its selection.
3. Selecting several compounds. This is done using the standard “mass” selection process. Mousing **h-s-left** when the mouse is not over any compound puts the system in selection mode. The mouse cursor changes into an upper left corner. Mousing **left** places the upper left corner of the selection box at the current mouse location. The mouse cursor now changes into a lower right corner. Mousing **left** places the lower right corner of the selection box at the current mouse location. Any compound whose display center falls within this box has its selection toggled.
4. Select compounds containing a particular substring. The **Compound String Search** command prompts the designer for a string. Any compound whose name contains this string is selected.

Selecting keywords is done by any of the first three methods described for selecting compounds.

The **Display Selected Choices** command creates a table containing values for each of the selected keywords for every selected compound. This table is displayed in the Data Display Pane. The **Create Correlation Matrix** command creates a simple correlation matrix between the selected keywords and displays it in the Data Display Pane. Both the table and correlation matrix are objects acceptable by other commands.

The **Create Function** commands allow functions of physical properties to be used as keywords. For example, we might wish to examine the linearity of  $T_c - T_b$  with  $Mw$ . We need to create a keyword which accesses the value  $T_c - T_b$ . The **Create Function** prompts for a name and value for a new keyword function. The system creates a **keyword-function** object which is added to the Keywords Pane. This new keyword is accessible in the same manner as any other keyword.

The **Plot Data** command displays two columns of a data display object in a 2D active graph. The data points plotted on 2D active graphs are objects themselves. This is an extremely desirable feature because this allows them to contain data and be accessed. The **Describe Data Point** and **Describe Data Points** commands enable the designer to describe any of the points displayed in a 2D active graph. The **Describe Bar** command enables the designer to list the compounds contained in any of the bars of a histogram. The Description Pane displays the results of accessing the information stored in these objects.

## 5.4 Data Configuration Objects

The seven major objects used in the Data Configuration are:

1. **data-base-molecule**
2. **data-object**
3. **data-base-keyword**
4. **keyword-function**
5. **data-point**
6. **data-display-object**
7. **correlation-matrix**

The definitions for these objects and their associated functions are in the file:

**molecular-design: data-base; objects.lisp.**

I discuss the important instance variables and functionality of each of these objects.

### 5.4.1 Data Base Molecule Object

The data base molecule object is the main data structure for storing physical property data. The important instance variables of the data base molecule object are shown in Table 5.2 with values for an example instance.

485 data base molecules are currently known to the system. These have all been defined using the **define-compound** macro. These definitions are in the file:

**molecular-design: data-base; physical-property-data.lisp.**

The values for the instance variables in Table 5.2 are enclosed in brackets, < ... >, to denote that these are not the actual values stored. What is actually stored in each variable is another object which contains these values. This object is described next.

Table 5.2: Data Base Molecule Instance Variables<sup>†</sup>

Instance Variable	Example Value
Formula	"CCl4"
Molecular Weight	< 153.823 ((g).(g-mol)) reid77 >
Critical Temperature	< 556.4 ((K)) ambrose80 >
Critical Pressure	< 45.0 ((atm)) ambrose80 >
Critical Volume	< 276.0 ((cc).(g-mol)) ambrose80 >
Critical Compressibility	< 0.272 ambrose80 >
Acentric Factor	< 0.193 ambrose80 >
Dipole Moment	< 0.0 ((debye)) ambrose80 >
Standard Gibbs Energy of Formation	< -13.92 ((kcal).(g-mol)) reid77 >
Standard Enthalpy of Formation	< -24.0 ((kcal).(g-mol)) reid77 >
Enthalpy of Vaporization at Tb	< 7170.0 ((cal).(g-mol)) reid77 >
Normal Freezing Point	< 250.0 ((K)) reid77 >
Normal Boiling Point	< 349.9 ((K)) ambrose80 >
Number of Atoms	< 5 >
Enthalpy of Fusion at Tm	< 601.0 ((cal).(g-mol)) stull69 >
Ideal Gas Heat Capacity	< 298.0 20.02 ((cal).(g-mol K)) stull69 >
	< 300.0 20.08 ((cal).(g-mol K)) stull69 >
	< 400.0 22.04 ((cal).(g-mol K)) stull69 >
	⋮
Liquid Viscosity	< 273.15 1.349 ((cp)) orrick73 >
	< 273.75 1.332 ((cp)) orrick73 >
	< 288.04 1.048 ((cp)) orrick73 >
	⋮
Factor 1	-
Factor 2	-
Factor 3	-

<sup>†</sup> Values from carbon tetrachloride data base molecule instance.

Table 5.3: Data Object Instance Variables<sup>†</sup>

Instance Variable	Example Value
Temperature	–
Pressure	–
Value	153.823
Units	((g).(g-mol))
Reference	reid77

<sup>†</sup> Values from molecular weight data object.

#### 5.4.2 Data Object

A data base molecule's instance variables need to contain more information than just the physical property's value. Additional information such as temperature and pressure are needed to fully specify the data point. This information is stored using a data object. Table 5.3 shows example values for the important instance variables of the data object. These values were taken from the molecular weight data object in Table 5.2.

All instance variables of a data base molecule store a list of data values. This enables temperature dependent properties like the ideal gas heat capacity and liquid viscosity to have their multiple values stored. Even if a physical property is not temperature or pressure dependent it is useful to be able to store several values from different references. Thus the normal boiling point instance variable of a data base molecule could contain several data object instances which differed in value and reference.

Each unit is represented by a collection of symbols using its standard names e.g., cal, BTU, K, F. The representation for a unit is a list of two lists. The first sublist contains all the unit symbols used in the numerator and the second sublist contains all

Table 5.4: Data Base References

Data Base	Bibliography
ambrose80	[1,2]
mcclellan63	[9]
orrick73	[13]
reid77	[14]
stull69	[17]

the units used in the denominator. Thus the representation of

$$\frac{cal}{g\text{-}mol \cdot K} \quad (5.1)$$

is

$$((cal) \quad (g\text{-}mol \ K)). \quad (5.2)$$

There are currently five references used in the data base. These are listed in Table 5.4 with their bibliographic references.

### 5.4.3 Data Base Keyword

The **data-base-keyword** object provides facilities to access the values of data object instances stored in data base molecules instance variables. It also provides the interface to the user for choosing which physical properties of a data base molecule are displayed.

The data base keyword object has three important instance variables. These are shown in Table 5.5 for an example instance.

The **variable-symbol** instance variable stores the symbol to be used in accessing the data object instances from the data base molecule. The pretty name is the string which is presented to the user in the Keywords Pane. The short name is a string which

Table 5.5: Data Base Keyword Instance Variables<sup>†</sup>

Instance Variable	Example Value
Variable Symbol	<code>molecular-weight</code>
Short Name	<code>Mw</code>
Pretty Name	<code>Molecular Weight</code>

<sup>†</sup> Molecular Weight Data Base Keyword

is used for the labeling of table columns and axes.

#### 5.4.4 Keyword Function Object

The keyword function object provides a facility for adding functions of data base keywords. For example, the reduced boiling point is not stored as an instance variable in a data base molecule. However, since the normal boiling point and the critical temperature are, the reduced boiling point can be formed by dividing the normal boiling point by the critical temperature. The important instance variables for the keyword function object are shown in Table 5.6 for an example instance.

The **label** instance variable stores the string which is presented to the user. As each keyword function is created its label is added to the Keywords Pane.

The **value** instance variable stores the LISP code which defines the function to

Table 5.6: Keyword Function Instance Variables:  $T_{br}$  Keyword Function

Instance Variable	Example Value
Label	<code>"Tbr"</code>
Value	<code>(/ normal-boiling-point critical-temperature)</code>
Needed Keywords	<code>(#&lt;data-base-keyword 1&gt; #&lt;data-base-keyword 2&gt;)</code>
Symbol List	<code>(normal-boiling-point critical-temperature)</code>

compute the value. Any of the existing data base keywords can be used in the creation of these functions. Other keywords functions can not be used in new functions.

The function stored in the value slot is parsed for data base keywords. These keyword instances are collected into a list and stored in the `needed-keywords` instance variable. This information is necessary for constructing the code which evaluates the function value.

The `symbol-list` instance variable stores a list of all the keyword symbols used in the function stored in the value instance variable. This list is used to identify the data base keywords used to construct the code which evaluates the function value.

#### 5.4.5 Data Display Object

The data display object is a two dimensional table displaying keyword and keyword function values for data base molecules. The display object is formed using all the selected data base molecules of the Compounds Pane, and all the selected data base keywords and keyword functions of the Keywords Pane. After a display object is created it is added to the top of the Data Display Pane. Figure 5.3 shows a `data-display-object` displayed in the Data Display Pane.

The data display object contains three important instance variables: 1) `compounds`; 2) `keywords`; 3) `data-array`. The `compounds` instance variable contains a list of all the data base molecules used to form this object. The `keywords` instance variable contains a list of all the data base keywords and keyword functions used to form this object. To simplify the code used to display the data display object, an array is formed which corresponds to the values retrieved by using the keywords to access values from the

### Data Base Section: Data Configuration

**Compounds**

- 3-Methyl-1-Butene, cis
- 3-Methyl-2-Pentene, trans
- 4-Methyl-2-Pentene, cis
- 4-Methyl-2-Pentene, trans
- 2,3-Dimethyl-1-Butene
- 2,3-Dimethyl-2-Butene
- 3,3-Dimethyl-1-Butene
- Cyclohexane
- Methyl Isobutyl Ketone
- n-Butyl Acetate
- Isobutyl Acetate
- Ethyl Butyrate
- Ethyl Isobutyrate

**Keywords**

- Formula
- Molecular weight
- Critical Temperature
- Critical Pressure
- Critical Volume
- Critical Compressibility
- Scatter Factor

**Data Display Pane**

Molecules	Formula	Mw	Tc	Pc
Argon	Ar	39.948	150.8	48.1
Boron Trichloride	BCl <sub>3</sub>	117.169	455.0	38.2
Boron Trifluoride	BF <sub>3</sub>	67.803	260.8	49.2
Bromine	Br <sub>2</sub>	159.808	568.0	10
Nitrosyl Chloride	ClNO	65.459	440.0	90
Chlorine	Cl <sub>2</sub>	70.906	415.9	78.7
Phosphorus Trichloride	Cl <sub>3</sub> P	137.333	563.0	--
Silicon Tetrachloride	Cl <sub>4</sub> Si	169.898	509.1	35.5
Deuterium	D <sub>2</sub>	4.032	38.2	16.3
Deuterium Oxide	D <sub>2</sub> O	20.031	644.0	213.8
Fluorine	F <sub>2</sub>	37.997	144.3	51.5
Nitrogen Trifluoride	F <sub>3</sub> N	71.002	234.0	44.7
Silicon Tetrafluoride	F <sub>4</sub> Si	104.08	259.0	36.7
Sulfur Hexafluoride	F <sub>6</sub> S	146.05	318.7	37.1
Hydrogen Bromide	HBr	80.912	363.2	84.4
Hydrogen Chloride	HC <sub>l</sub>	36.461	324.7	92
Hydrogen Fluoride	HF	20.006	451.0	54
Hydrogen Iodide	HI	127.912	424.0	82
Hydrogen	H <sub>2</sub>	2.016	33.6	12.9
Water	H <sub>2</sub> O	18.015	647.1	217.6
Hydrogen Sulfide	H <sub>2</sub> S	34.08	373.2	88.2
Ammonia	H <sub>3</sub> N	17.031	405.5	111.3
Hydrazine	H <sub>4</sub> N <sub>2</sub>	32.045	653.0	14
Helium-4	He	4.003	5.19	2.24

MD Command:

- Display Data-Base Data Statistics
- Display Functions
- Display Selected Choices
- Form Correlation Matrix
- Locate Compound

Figure 5.3: Display of Physical Property Data in Tabular Format

compounds. This array is stored in the `data-array` instance variable. The data array is formed after the creation of the instance. Display of the data display object thus loops through the array presenting each element. This is much quicker than having to access each of the molecules each time the display object is represented.

Data display objects are used as the interface to the graphical display of physical property data. The two graphical display objects, 2D active graphs and histograms, both take input via a data display object.

#### **5.4.6 Correlation Matrix Object**

The correlation matrix object displays a matrix showing the correlations among the selected data base keywords and keyword functions computed from the set of selected data base molecules. After a correlation matrix object is created it is added to the top of the Data Display Pane. Figure 5.4 shows a screen of the system displaying a correlation matrix.

### **5.5 Data Configuration Commands**

The following commands are listed in the Data Configuration's Command Menu. The definitions of these commands are in the file:

```
molecular-design:base;commands.lisp.
```

**Clear Data Display:** Removes all the data display objects and correlation matrices from the Data Display Pane.

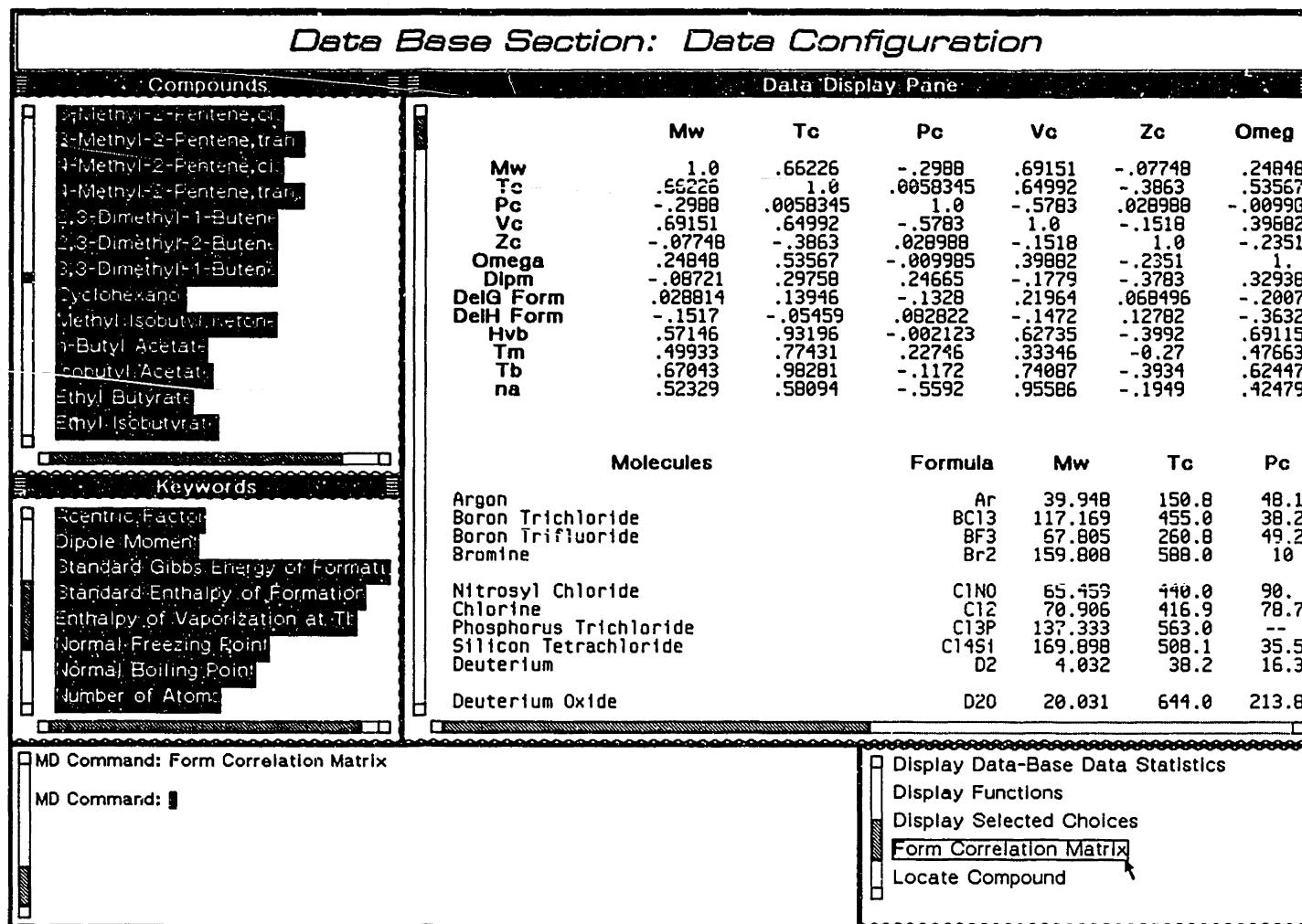


Figure 5.4: Correlation Matrix Display

**Compound String Search:** Prompts for a substring to be searched for in a compound's name. All compounds whose names contain this substring are selected in the Compounds Pane.

**Create Histogram:** Creates a histogram of one of the variables displayed in a data display object. This command prompts for a data display array. The possible variables for which a histogram can be formed are displayed for choice by the user. Selecting a variable creates a histogram, changes to the Data-Base Section: Plot Configuration, and adds the histogram to the Plot Pane.

**Create Keyword Function:** Creates a function which is added to the keywords displayed in the Keywords Pane. This command prompts for the label which will be displayed in the Keywords Pane and the function value used to generate the new data points. I did not implement code to save these functions to a file.

**Deselect All Compounds:** Deselects all the selected compounds displayed in the Compounds Pane. Selected compounds are used to form **data-display** and **correlation-matrix** objects.

**Deselect All Keywords:** Deselects all the selected keywords displayed in the Keywords Pane. Selected keywords are used to form **data-display** and **correlation-matrix** objects.

**Display Data-Base Data Statistics:** Uses a window resource exposed over the Data Display Pane to display information on the status of objects in the configuration.

Statistics displayed are:

1. Number of Compounds Present.
2. Number of Compounds Selected.
3. Number of Keywords Present.
4. Number of Keywords Selected.
5. Number of Correlation Matrices Displayed.
6. Number of Data-Matrices Displayed.

**Display Functions:** Uses a window resource exposed over the Data Display Pane to display the user defined keyword functions. The label and function value are displayed for all the keyword functions known to the system.

**Display Selected Choices:** This command creates a data display object. The selected data base molecules of the Compounds Pane and the selected data base keywords and keyword functions of the Keywords Pane are used to form the data display object. After the object is created it is added to the beginning of the Data Display Pane.

**Locate Compound:** Prompts for a string which is searched for in a compound's name. The Compounds Pane scrolls so that the first compound whose name contains this string is displayed at the top of the window.

**Plot Data:** Creates a 2D active graph. The data to be plotted in this graph must be contained in a displayed data display object. The command prompts for a data display object. The keywords of the data display object are then presented in two lines requesting selection of the x and y axes keywords. For example, the data display object shown in Figure 5.3 has the keywords:

Formula	Mw	Tc	Pc	omega	Tb.
---------	----	----	----	-------	-----

The prompt shown to the designer is:

**X Data:** Mw Tc Pc omega Tb  
**Y Data:** Mw Tc Pc omega Tb

The designer then mouses on one keyword for the x data and one keyword for the y data. Once these are chosen the data is extracted and the 2D active graph is formed. This new graph is added to the Plot Configuration's Plot Pane.

**Select all Compounds:** Selects all the unselected compounds displayed in the Compounds Pane. Selected compounds are used to form **data-display** and **correlation-matrix** objects.

**Select all Keywords:** Selects all the unselected keywords displayed in the Keywords Pane. Selected keywords are used to form **data-display** and **correlation-matrix** objects.

## 5.6 Plot Configuration Objects

The five major objects used in the Plot Configuration are:

1. **2d-active-graph**
2. **2d-data-point**
3. **function-graphic**
4. **histogram**

The definitions of the 2D data point object and some of its associated functions are in the file:

`molecular-design:base;objects.lisp.`

The definitions of the other objects and their associated functions are in the file:

`molecular-design:active-graphs;objects.lisp.`

I discuss the important instance variables and functionality of each of these objects.

### 5.6.1 2D Active Graph Object

The 2D Active Graph object exemplifies the concepts of object oriented programming.

The task of a 2D Active Graph is to display data and functions Figure 5.5 shows a screen of the configuration with a typical 2D active graph displayed in the Plot Pane.

The `2-D-active-graph` object contains over 30 instance variables. The majority of these store information needed to present the graph e.g., character styles, tick lengths, and label and title strings. The names associated with most of these instance variables is sufficient to describe their purpose. Two instance variables:

1. `displayed-points`
2. `displayed-functions`

require further explanation.

The `displayed-points` instance variable contains a list of `data-points` currently displayed in the graph. The `displayed-functions` instance variable contains a list of `graphic-functions` currently displayed in the graph. Both `data-points` and `graphic-functions` are objects with generic functions specifying their own display. Whenever a `2-D-active-graph` object is presented the system loops through the lists in the `displayed-points` and `displayed-functions` instance variables presenting each of the objects.

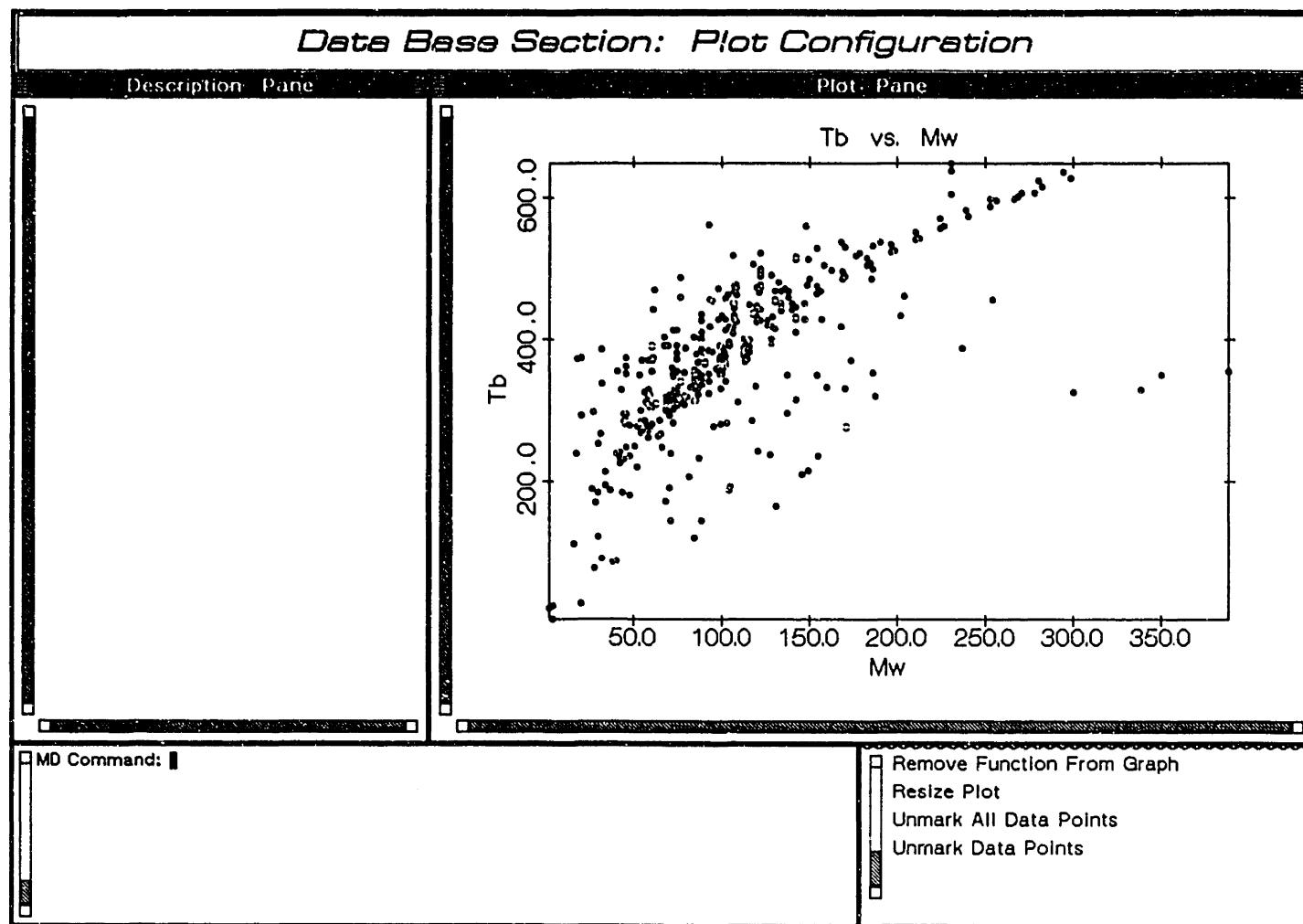


Figure 5.5: Plot Configuration with 2D Active Graph

## 5.6.2 2D Data Point Object

Every data point contains the molecule object from which its x and y values were obtained. This gives the possibility of answering queries about other properties not currently displayed in the active graph. A 2-D-point object stores its coordinates in the instance variables **x-coordinate** and **y-coordinate**.

The **shape** instance variable specifies the graphic image used to display the data point. There are eight possible images:

Point	Circle	Filled Circle	Triangle
Filled Triangle	Square	Filled Square	Cross

The **metric** instance variable stores a number used to size the image. This corresponds to the radius of a circle or the half-side length of a polygon.

## 5.6.3 Function Graphic Object

The **function-graphic** object is used to graphically display functions on 2-D-active-graphs. The object can only display functions which are explicit in one of the variables. The **graph** instance variable of the object stores the 2-D-active-graph object on which the function is displayed. The system uses the upper and lower limits of this graph to calculate a range of function values. These values are used to draw the function.

## 5.6.4 Histogram Object

Like the 2D Active Graph, the histogram object displays data graphically. The histogram is constructed similarly to the 2D Active Graph in that a Data Display Object

is first chosen. The designer is presented with the columns the Data Display Object contains. Choosing a column causes the system to access all the data from that column of the Data Display Object and use it to form the histogram. Figure 5.6 shows a screen of the Plot Configuration with a histogram displayed.

Analogous to the 2D Active Graphs the objects contained within the graph are objects themselves and thus contain information. The only object contained within the histogram are bars. Each bar contains the list of data points used to determine its size and location. The Describe Bar command displays the data points which compose the bar in the Description Pane. Figure 5.6 shows the Description Pane displaying the contents of the rightmost bar of the histogram.

Editing facilities provide for reassignment of the number of categories into which the allocation is made.

## 5.7 Plot Configuration Commands

The following commands are listed in the Plot Configuration's Command Menu. The definitions of these commands are in the file:

```
molecular-design:database;commands.lisp.
```

**Clear Description Pane:** Removes all the descriptions displayed in the Description Pane.

**Clear Plot Pane:** Removes all the 2-D-active-graphs and histograms from the Plot Pane. Once these objects are deleted they can not be retrieved.

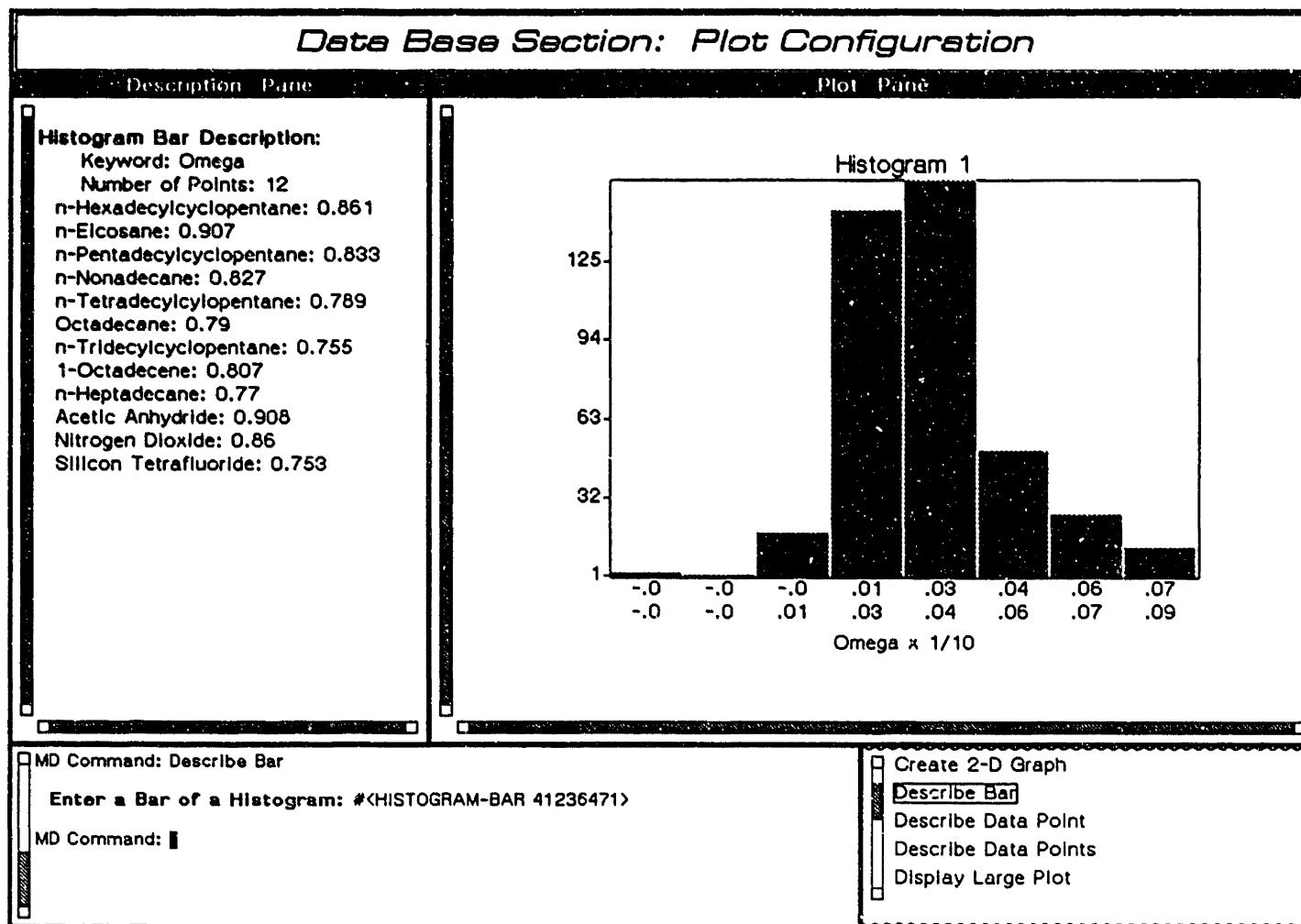


Figure 5.6: Acentric Factor Histogram

**Describe Bar:** Prompts for a histogram bar. Displays the physical property value in the Description Pane for each of the compounds contained in that bar.

**Describe Data Point:** Prompts the designer to mouse on a single data point. The name of the compound and its values for the x and y physical properties are displayed in the Description Pane.

**Describe Data Points:** Prompts the designer to enclose a set of data points using the “rubber-banding box” procedure. The names and x and y physical property values for each of the data points enclosed by the box are displayed in the Description Pane. Figure 5.7 shows an example of describing several data points.

**Display Large Plot:** This command displays a chosen 2D active graph in a window the size of the screen. The purpose of this large display was to improve the quality of bitmaps made of the system.

**Display Statistics:** Prompts for a Histogram. Various statistics for the chosen histogram are displayed in a window resource exposed over the Plot Pane. Figure 5.8 shows an example of this display. The median value is obviously incorrect.

**Fit Line To Points:** Prompts the designer for a 2D active graph. A line is regressed through all the unmarked data points of the chosen graph. The line is displayed graphically on the graph and its equation is displayed in the Molecular Design Interaction Pane.

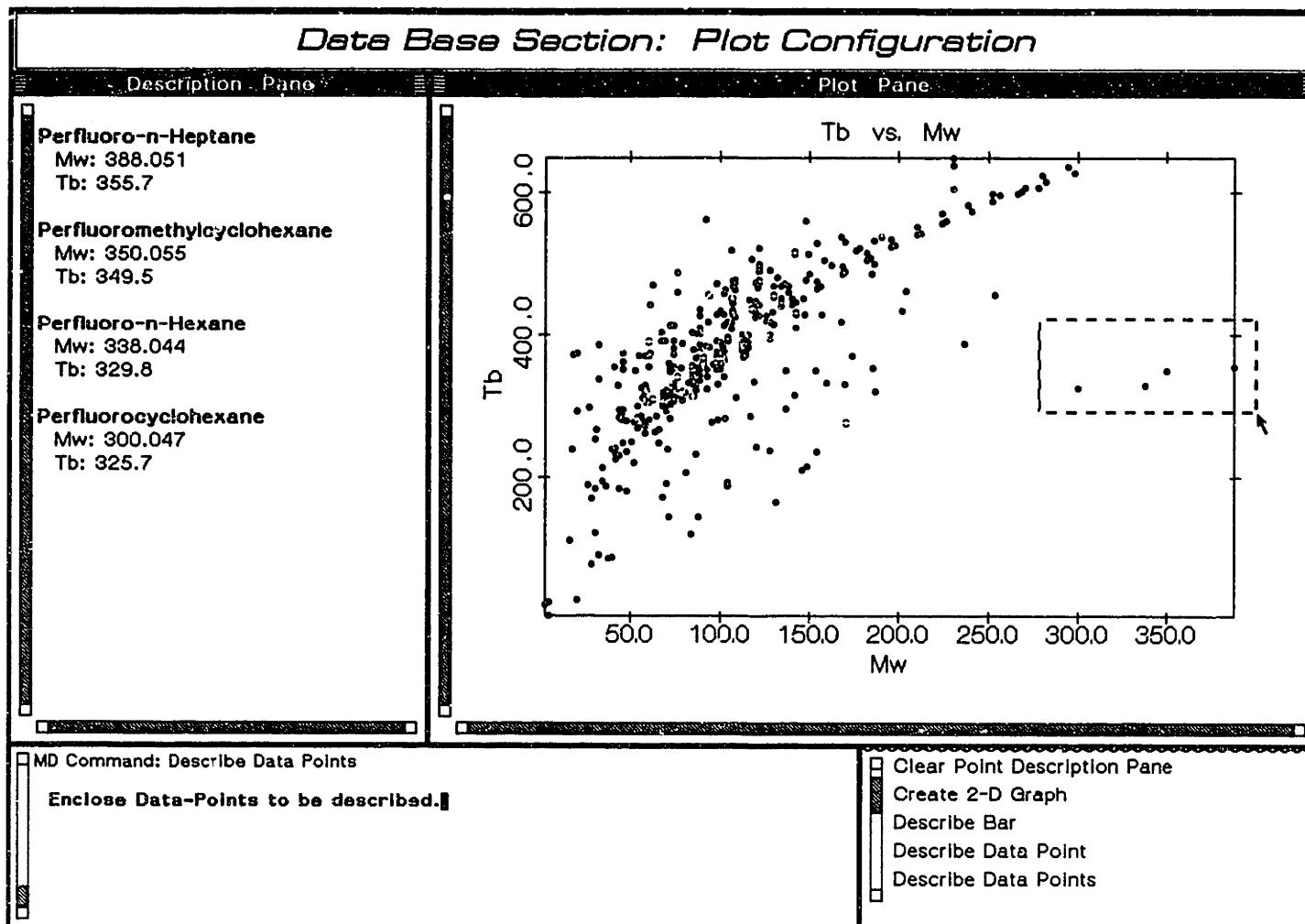


Figure 5.7: Describing Several Data Points Displayed in a 2D Active Graph

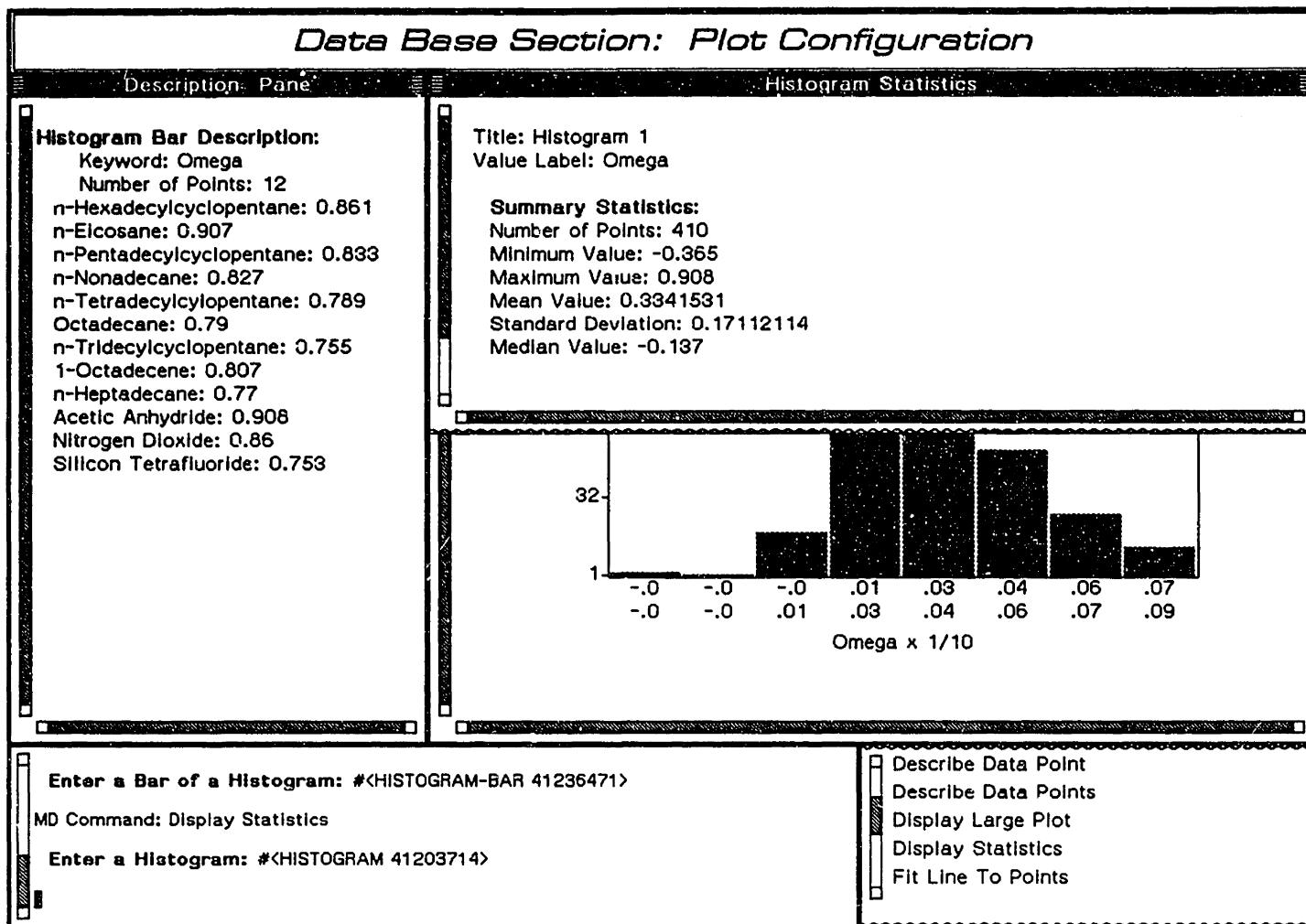


Figure 5.8: Histogram Statistics Display

**Fit Polynomial To Points:** First prompts the designer for a 2D active graph. The order of the polynomial is then prompted for. A polynomial is regressed through all the unmarked data points of the chosen graph. The polynomial is displayed graphically on the graph and its equation is displayed in the Molecular Design Interaction Pane.

**Mark Data Points:** Prompts the designer to enclose a set of data points using the “rubber-banding box” procedure. The enclosed points are “marked”. Their display changes from a filled circle to a cross. Marked points are not included in linear or polynomial fits. Marking points thus represent one way in which outliers can be excluded from a regression.

Figure 5.9 shows the effect of marking points. Two quadratics were fit to  $T_b$  vs.  $M_w$  data. The first quadratic which decreases rapidly at high molecular weights was regressed using all the data. The second quadratic was regressed with the four perfluorocarbons marked and thus excluded from the regression.

**Remove All Functions from Graph:** Prompts the designer for a 2D active graph. All the graphic functions displayed in this graph are removed.

**Remove Function From Graph:** Prompts the designer for a displayed graphic function. Graphic functions are created with either the **Fit Line To Points** or **Fit Polynomial To Points** command. The chosen function is removed from the graph.

**Resize Plot:** Prompts the designer for a 2D active graph. The new size of this graph is specified with the mouse via the “rubber-banding box” procedure.

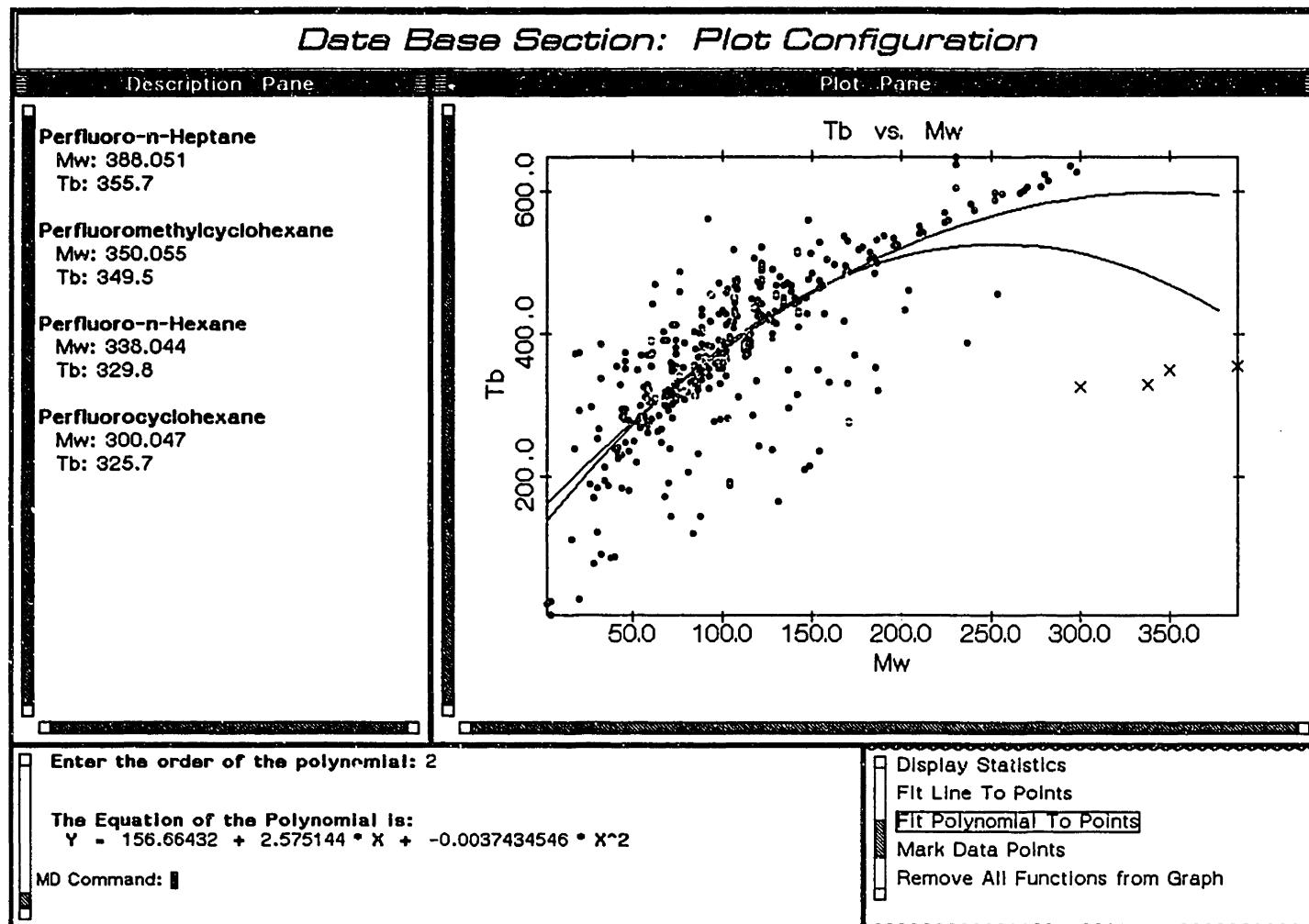


Figure 5.9: Polynomial regressions. Using marked points to exclude data.

**Unmark All Data Points:** Prompts the designer for a 2D active graph. All the marked data points in the chosen graph are unmarked.

**Unmark Data Points:** Prompts the designer to enclose a set of data points using the “rubber-banding box” procedure. The enclosed points are unmarked.

## 5.8 Section Discussion

Besides storing data for use in problem formulation and estimation technique evaluation the Data Base Section represents initial facilities for online development of estimation techniques. New estimation techniques could be developed specifically tailored for the design problem being addressed. Currently estimation techniques are developed for general use. They attempt to estimate the ethane's property as well as dodecane's property. If a designer was searching for new refrigerants then estimation techniques which estimated C1–C5s would suffice. Restricting the scope of estimation techniques would improve their accuracy.

The following facilities need to be added to the Data Base Section to develop new estimation techniques:

1. Multiple regression analysis.
2. Principal component analysis.
3. Optimization analysis.
4. Factor analysis.

## 5.9 Example Usage

These instructions detail a step-by-step example usage of the Data Base Section. We accomplish the following tasks using the facilities of both the Data Configuration and Plot Configuration:

1. Display physical property data in a **data-display-object**.
2. Create a **correlation-matrix**.
3. Display two columns of a **data-display-object** in a **2D-active-graph**.
4. Create a **histogram** showing the distribution of physical property data.
5. Regress polynomial equations to data.

### Move to the Data Configuration

If the current configuration is not the Data-Base Section: Data Configuration the first task is to change configurations.

**Action 5.1** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 5.2** *Mouse left on the Data-Base Section Data Configuration.*

The system changes to the Data Configuration.

### Tabular Data Display

The Data Configuration is for the tabular display of physical property data. The major object used in this tabular display is the **data-display-object**. We form a **data-dis-**

**play-object** by first selecting several compounds and keywords. In this example we select all the displayed compounds. Six keywords are selected:

- |                         |                         |
|-------------------------|-------------------------|
| 1) Formula              | 4) Critical Pressure    |
| 2) Molecular Weight     | 5) Acentric Factor      |
| 3) Critical Temperature | 6) Normal Boiling Point |

After these objects are selected the **Display Selected Choices** command creates and displays the **data-display-object** in the Data Display Pane.

**Action 5.3** *Mouse left on the Select All Compounds command.*

**Action 5.4** *Mouse h-sh-left on the Formula keyword displayed in the Keywords Pane.*

**Action 5.5** *Mouse h-sh-left on the Molecular Weight keyword displayed in the Keywords Pane.*

**Action 5.6** *Mouse h-sh-left on the Critical Temperature keyword displayed in the Keywords Pane.*

**Action 5.7** *Mouse h-sh-left on the Critical Pressure keyword displayed in the Keywords Pane.*

**Action 5.8** *Mouse h-sh-left on the Acentric Factor keyword displayed in the Keywords Pane.*

**Action 5.9** *Mouse h-sh-left on the Normal Boiling Point keyword displayed in the Keywords Pane.*

**Action 5.10** *Mouse left on the Display Selected Choices command.*

The display of physical property data for all 485 compounds may take several minutes.

The screen display should be similar to Figure 5.3.

### Graphical Data Display

Displaying data in a tabular format is the first step in displaying physical property data graphically.

**Action 5.11** *Mouse left on the Plot Data command.*

The system prompts for a data-display-object:

**Enter a Data Display Array:**

**Action 5.12** *Mouse left on the data-display-object currently displayed in the Data Display Pane.*

The keywords of the data table are displayed in the Molecular Design Interaction Pane:

**X Data:** Mw Tc Pc Omega Tb

**Y Data:** Mw Tc Pc Omega Tb

**ABORT** aborts, **END** uses these values

**Action 5.13** *Mouse left on Mw in the X Data: line.*

**Action 5.14** *Mouse left on Tb in the Y Data: line.*

**Action 5.15** *Press the <END> key.*

The system changes to the Plot Configuration. A 2D-active-graph is formed and displayed in the Plot Pane. The plot defaults to a size which does not occupy the entire pane. We resize the plot to make it easier to see.

**Action 5.16** *Mouse left on the Resize Plot command.*

The system prompts for a plot to be resized:

**Enter a plot to be resized:**

**Action 5.17** *Mouse left on the 2D-active-graph displayed in the Plot Pane.*

The mouse cursor moves to the Plot Pane and changes into an upperleft corner.

**Action 5.18** *Mouse left near the upper left corner of the Plot Pane.*

This positions the new upper left corner of the plot. The mouse cursor now appears as a lower right corner. As you move the mouse a “rubber-banding box” is drawn connecting the affixed upper left corner and the lower right corner mouse cursor. This box roughly shows the new size of the plot.

**Action 5.19** *Mouse left near the lower right corner of the Plot Pane.*

## Describing Data Points

Each of the data points displayed in the 2D-active-graph is a data-point object. Each data point is presented on the screen and is thus accessible.

**Action 5.20** *Mouse left on the Describe Data Points command.*

The mouse cursor changes to an upper left corner.

**Action 5.21** *Position the mouse cursor to the upper left of several points you wished described. Mouse left when the cursor is in place.*

The upper left corner is affixed to the mouse cursor location. The mouse cursor then changes to a lower right corner. This lower right corner is connected to the upper left corner via a “rubber-banding box”. The data points enclosed by this box are described.

**Action 5.22** *Enclose the data points to be described with the rubber-banding box.*

**Mouse left when the cursor is in position.**

The compounds the data points represent and their exact X and Y coordinates are displayed in the Data Description Pane.

## Fitting Functions

The Data-Base Section provides linear and polynomial regressions.

**Action 5.23** *Mouse left on the Fit Polynomial To Points command.*

The system prompts for the plot whose data is to be regressed:

**Click on the Plot whose data is to be fitted.**

**Action 5.24** *Mouse left on the 2D-active-graph displayed in the Plot Pane.*

The system next prompts for the order of the polynomial fit:

**Enter the order of the polynomial:**

**Action 5.25** *Enter the number 2 and press return.*

The system collects all the data points and fits a second order polynomial to them. The function is displayed on the plot and the equation is displayed in the Molecular Design Interaction Pane:

**The Equation of the Polynomial is:**

$$Y = 131.16403 + 3.1156569 * X + -0.0061297873 * X^2$$

The polynomial fit is greatly effected by the four outliers at high molecular weights but moderate normal boiling points. Marking these data points removes them from consideration when creating the polynomial.

**Action 5.26** *Mouse left on the Mark Data Points command.*

Again the system uses a “rubber-banding box” to choose the data points.

**Action 5.27** *Position the mouse cursor above and to the left of the four outliers.*  
*Mouse left when the cursor is positioned.*

The mouse cursor changes to a lower right corner connected to the affixed upper left corner by a rubber-banding box.

**Action 5.28** *Position the mouse cursor below and to the right of the four outliers.*  
*Mouse left when the cursor is positioned.*

The data points representing the four outliers change their presentation from four filled circles to four crosses. These “marked” data points are not included in further regressions.

**Action 5.29** *Mouse left on the Fit Polynomial To Points command.*

The system prompts for the plot whose data is to be regressed:

**Click on the Plot whose data is to be fitted.**

**Action 5.30** *Mouse left on the 2D-active-graph displayed in the Plot Pane.*

The system next prompts for the order of the polynomial:

**Enter the order of the polynomial:**

**Action 5.31** *Enter the number 2 and press return.*

The system collects all the data points except the four marked outliers and fits a second order polynomial to them. The function is displayed on the plot and the equation is displayed in the Molecular Design Interaction Pane:

**The Equation of the Polynomial is:**

$$Y = 156.66432 + 2.575144 * X + -0.0037434546 * X^2$$

The new polynomial better approximates the data at high molecular weight once outliers are removed.

## **Creating Histograms**

The system provides facilities for creating and examining histograms of physical property data. Histograms are also displayed in the Plot Pane. Before creating a histogram we clean up the Plot Configuration.

**Action 5.32** *Mouse left on the Clear Plot Pane command.*

**Action 5.33** *Mouse left on the Clear Point Description Pane command.*

We now change configurations back to the Data Configuration.

**Action 5.34** *Type into the Molecular Design Interaction Pane the command: Previous Section.*

The system changes to the Data Configuration. The **data-display-object** we previously created is still displayed. Changing configurations causes the Data Description Pane to be redisplayed. Redisplaying the **data-display-object** takes about 3 minutes.

Once the Data Configuration is completely redisplayed we begin creating the histogram.

**Action 5.35** *Mouse left on the Create Histogram command.*

The system prompts for a **data-display-object**:

**Enter a Data Display Array:**

**Action 5.36** *Mouse left on the **data-display-object** displayed in the Data Display Pane.*

The system displays keywords of the **data-display-object**.

**Choose Column for Display:** Mw Tc Pc Omega Tb

**ABORT** aborts, **END** uses these values

**Action 5.37** *Mouse left on the Omega keyword.*

**Action 5.38** *Press the <END> key.*

The system changes to the Plot Configuration. A histogram is displayed in the Plot Pane. We resize the histogram for better appearance.

**Action 5.39** *Mouse left on the Resize Plot command.*

Again the system uses the “rubber-banding box” to accept the new size. The system prompts for a plot to be resized:

**Enter a plot to be resized:**

**Action 5.40** *Mouse left on the histogram displayed in the Plot Pane.*

The mouse cursor moves to the Plot Pane and changes into an upperleft corner.

**Action 5.41** *Mouse left near the upper left corner of the Plot Pane.*

This positions the new upper left corner of the histogram. The mouse cursor now appears as a lower right corner. As you move the mouse a “rubber-banding box” is drawn connecting the affixed upper left corner and the lower right corner mouse cursor. This box roughly shows the histogram’s new size.

**Action 5.42** *Mouse left near the lower right corner of the Plot Pane.*

## Histogram Descriptions

Analogous to describing a data point in a plot, the bars of a histogram can also be described.

**Action 5.43** *Mouse left on the Describe Bar command.*

The system prompts for the bar of a histogram:

**Enter a Bar of a Histogram:**

**Action 5.44** *Mouse left on the rightmost bar of the histogram displayed in the Plot Pane.*

The “contents” of the histogram bar are displayed in the Point Description Pane.

Summary statistics are available for histograms. These statistics include:

1. Number of Points
2. Minimum Value
3. Maximum Value
4. Mean Value
5. Standard Deviation
6. Median Value

**Action 5.45** *Mouse left on the Display Statistics command.*

The system prompts for a histogram:

**Enter a Histogram:**

**Action 5.46** *Mouse left on the displayed histogram.*

Statistics are displayed in a window resource exposed over the Plot Pane. Figure 5.8 showed an example display.

## Correlation Matrix

The Data Configuration provides a facility for creating simple correlation matrices.

**Action 5.47** *Type into the Molecular Design Interaction Pane the command: Previous Section.*

The system changes to the Data Configuration. We select additional keywords used to form the correlation matrix.

**Action 5.48** *Mouse h-sh-left on the Critical Volume keyword.*

**Action 5.49** *Mouse h-sh-left on the Critical Compressibility keyword.*

**Action 5.50** *Mouse h-sh-left on the Dipole Moment keyword.*

**Action 5.51** *Mouse h-sh-left on the Standard Gibbs Energy of Formation keyword.*

**Action 5.52** *Mouse h-sh-left on the Standard Enthalpy of Formation keyword.*

**Action 5.53** *Mouse h-sh-left on the Enthalpy of Vaporization at Tb keyword.*

**Action 5.54** *Mouse h-sh-left on the Normal Freezing Point keyword.*

**Action 5.55** *Mouse h-sh-left on the Number of Atoms keyword.*

Correlation matrices are displayed in the Data Display Pane. To speed redisplay of the pane first remove the data-display-object.

**Action 5.56** *Mouse left on the Clear Data Display Pane command.*

Now create and display the correlation matrix.

**Action 5.57** *Mouse left on the Form Correlation Matrix command.*

# Chapter 6

## Group Contribution Section

Physical property estimation procedures are the heart of my design procedures. The Group Contribution Section provides facilities for entering group contribution and equation oriented techniques. Models in the form of LISP code are entered for both types of estimation techniques. Additionally, groups and their contributions are specified for group contribution techniques.

The section is divided into two configurations: 1) Editing Configuration; 2) Model Entry Configuration. The editing configuration provides facilities for creating new groups. The designer connects atoms with bonds to form new groups. The model entry configuration provides facilities for entering contributions for these groups, models for group contribution techniques, and models for equation oriented estimation techniques.

### 6.1 Editing Configuration Layout

The screen layout of the editing configuration is shown in Figure 6.1. The screen real

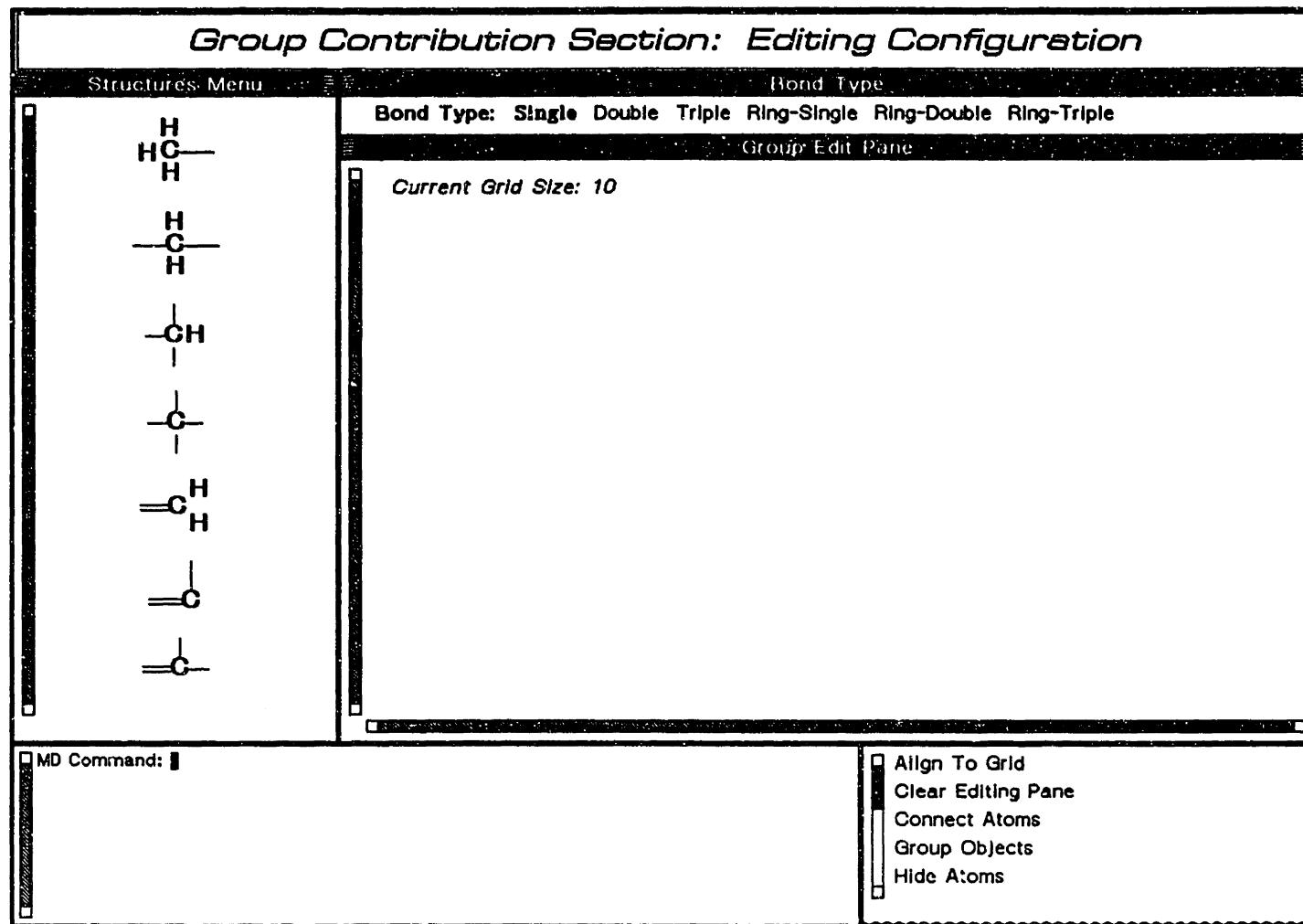


Figure 6.1: Group Contribution Editing Configuration Screen

estate is used by six panes:

**Group Contribution Section Editing Configuration Title Pane:** Displays the title of the Editing Configuration.

**Structures Menu:** Displays a list of all the groups known to the system. As new groups are created they are added to this pane.

**Bond Type Pane:** This is an AVV-pane which displays the “current” bond type. As atoms are connected with bonds the bond type used is determined from the value in this pane.

**Group Edit Pane:** This pane is the major focus of the Editing Configuration. The designer selects atoms which are entered into this pane. The pane acts as a canvas on which the designer arranges and connects atoms together to form groups. Groups constructed in this pane can be saved to file and made available to all other sections of the system.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Group Contribution Section Editing Configuration Commands Menu:** The command menu containing commands relevant to the Editing Configuration.

## 6.2 Model Entry Configuration Layout

The screen layout of the Model Entry Configuration is shown in Figure 6.2. The screen real estate is used by six panes:

**Group Contribution Section Model Entry Configuration Title Pane:** Displays the title of the Model Entry Configuration.

**Chosen Groups Pane:** Displays the groups which have had contributions specified. When developing a group contribution estimation technique the designer chooses a group and specifies its contribution. After specification each group is displayed in this pane with its specified physical property.

**All Groups Pane:** Displays all the groups known to the system.

**Technique Description Pane:** This pane displays estimation technique objects. Estimation technique objects contain much of the information needed to construct a new estimation technique.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Group Contribution Section Model Entry Configuration Commands Menu:** The command menu containing commands relevant to the Model Entry Configuration.

## Group Contribution Section: Model Entry Configuration

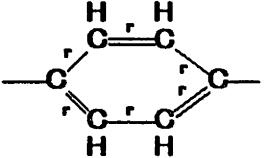
All Groups	Chosen Groups	Technique Description
		
		
		
		
		
		
<b>Entry Page</b>		
<input type="checkbox"/> MD Command: <input type="text" value=""/>		<input type="checkbox"/> Remove All Chosen Groups <input type="checkbox"/> Save Technique <input type="checkbox"/> Show All Groups <input type="checkbox"/> Specify Input Groups

Figure 6.2: Group Contribution Model Entry Configuration Screen

## 6.3 Section Operation

The Group Contribution Section addresses the task of entering new estimation techniques into the system. This may necessitate creating new groups. Entering parameters and models for estimation techniques is done in the Model Entry Configuration. Creating new groups is done in the Editing Configuration.

### 6.3.1 Creating Estimation Techniques

The **Create Property** command of the Problem Formulation Section and the **Create Estimation Technique** command of the Model Entry Configuration initiate the development of new estimation techniques. The **technique-input-object** object stores the description of the estimation technique while the designer inputs the necessary information. This information is prompted for by a menu when either of the **Create Property** or **Create Estimation Technique** commands are activated. The menu requests the following information:

**Type:** Group Contribution or Equation Oriented.

**Name:** A string the estimation technique is called.

**Class:** An estimation technique class. Estimation techniques are typically displayed arranged by estimation technique class.

**Estimated Property:** The physical property this technique estimates.

**State Variables:** A list containing *Temperature*, *Pressure*, or both.

**Required Properties:** Applicable for equation oriented techniques only. The physical properties required to determine the estimated property.

The input information is stored into a **technique-input-object** which is added to the Technique Description Pane.

Models used to estimate the physical property must be entered for both group contribution and equation oriented techniques. The **Edit Estimation Model** and **Enter Interval Model** commands activate the Entry Pane editor. The Entry Pane is a ZMACS style editor. The designer enters the models having access to most ZMACS commands. Typing <END> stores the entered code in the **technique-input-object**.

Group contribution techniques require contributions for a set of groups. All the groups known to the system are displayed in the All Groups Pane. The **Choose Group** and **Choose Groups** commands both prompt the designer to choose a group displayed in the All Groups Pane and then specify a contribution for it. The chosen group is displayed in the Chosen Groups Pane with its associated contribution displayed to the lower right. Figure 6.3 shows the Model Entry Configuration with a group contribution technique being developed. The designer has specified the contributions for several groups. These are displayed in the Chosen Groups Pane.

Once the input information, models, and group contributions are entered the technique is ready to be activated or saved. The **Activate Technique** command takes the information stored in the **technique-input-object** and creates an estimation technique. This technique is accessible to the system but is not stored in a file. The **Save Technique** command saves the information stored in the **technique-input-object** into a file. This technique is accessible the next time the system is compiled.

### 6.3.2 Creating New Groups

New group contribution techniques may use groups which are not currently known to the system. The primary task of the Editing Configuration is to create new groups.

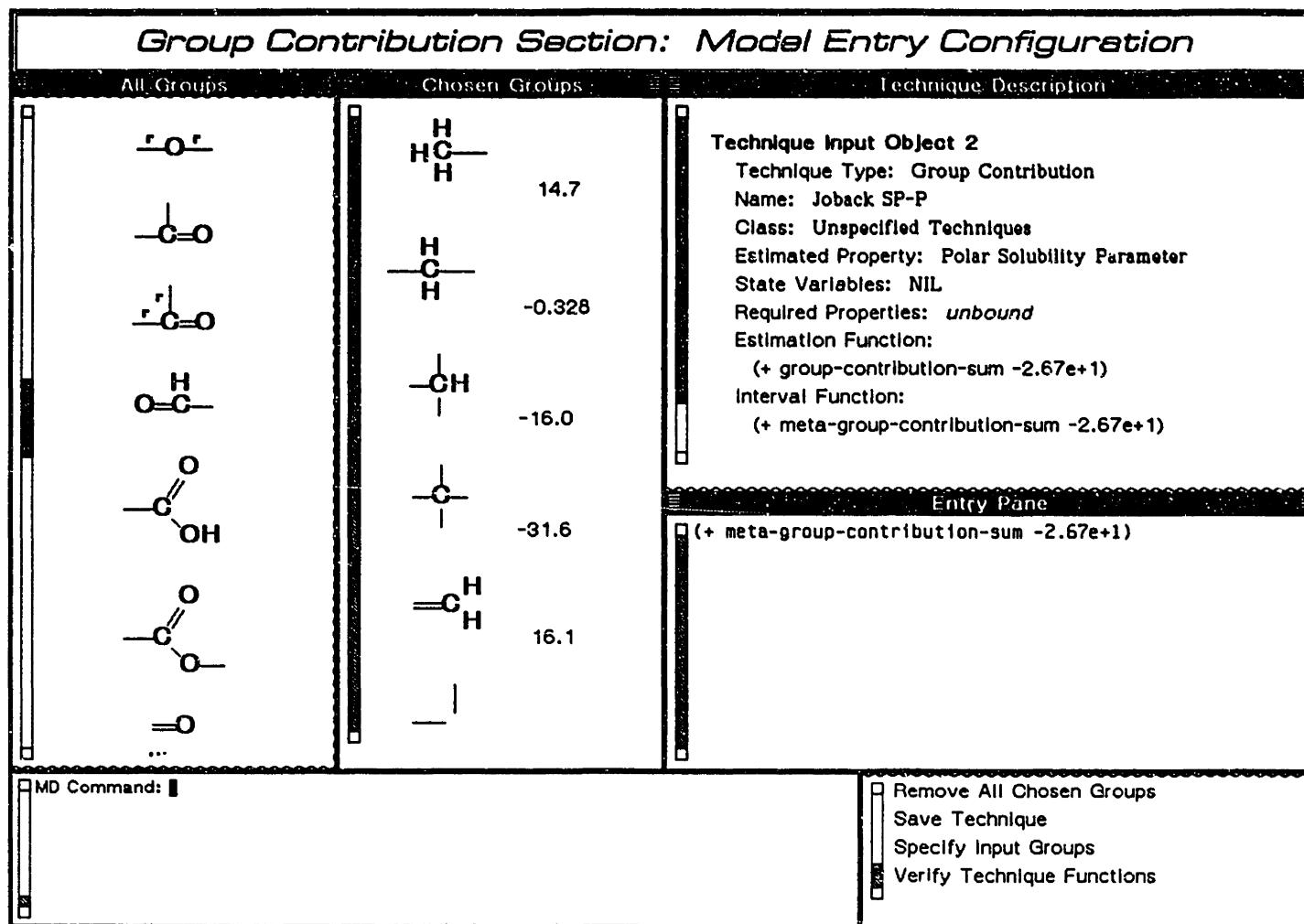


Figure 6.3: Entry of a Group Contribution Estimation Technique

The **Input Atoms** command prompts the designer for a sequence of atoms. These atoms are entered into the Group Edit Pane. The designer arranges these atoms into an aesthetically pleasing configuration. The **Connect Atoms** command places the designer into a loop in which consecutive mouse clicks on atoms connect them with a bond. The type of this bond is specified by the current bond type displayed in the Bond Type Pane.

The **Group Objects** command collects all the selected atoms and bonds in the Group Edit Pane and creates a new group. This group object is placed in the Structures Menu. The **Mass Select Atoms**, **Mass Select Bonds**, **Mass Deselect Atoms**, and **Mass Deselect Bonds** commands manipulate the selection of atoms and bonds. The **Hide Atoms** command allows atoms to be hidden from view. Hiding atoms sometimes produces a more aesthetically pleasing group. Hiding hydrogens in cyclic hydrocarbons is an example.

## 6.4 Editing Configuration Objects

The three major objects used in the editing configuration are:

1. **md-atom**
2. **bond**
3. **group**

The definitions of these objects and their associated functions are in the file:

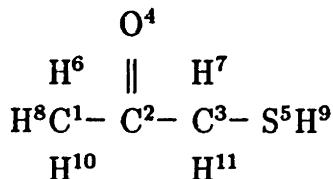
`molecular-design:group-contribution-section;objects.lisp.`

Groups are central to both the interactive and automatic design procedures. The representation chosen for storing, retrieving, and manipulating groups is thus important. There are several reviews of computer-compatible structure representations[8, 10,11,12]. The National Academy of Sciences classified representations into three categories[12]:

1. Linear Notations
2. Tabular and Graphic Representations
3. Fragment Codes

Linear notations make entry of molecular structure via a keyboard very easy. They are also particularly suitable in information retrieval systems involving large data bases of chemical structures[6]. Wiswesser linear notation(WLN)[15] is a popular linear notational system that represents structures with character strings. The notation uses a set of complex *ad hoc* rules to derive names for structures. Most chemists are incapable of deriving structural diagrams from such names, and vice versa. Computer programs for translation of structural diagrams into names have been only partially successful, primarily because of the complexity of the rules[6].

Connection matrices, tables, and atom-connectivity lists provide the most valuable representation for the vast majority of applications in chemical structure elucidation and manipulation[6]. These methods capture topology in tabular forms that directly represent structural connectivity. In the simplest case a molecule is represented by the connections between atoms. The connection table for:



is given by the following matrix:

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	O <sub>4</sub>	S <sub>5</sub>	H <sub>6</sub>	H <sub>7</sub>	H <sub>8</sub>	H <sub>9</sub>	H <sub>10</sub>	H <sub>11</sub>
C <sub>1</sub>	X	S	-	-	-	S	-	S	-	S	-
C <sub>2</sub>	S	X	S	D	-	-	-	-	-	-	-
C <sub>3</sub>	-	S	X	-	S	-	S	-	-	-	S
O <sub>4</sub>	-	D	-	X	-	-	-	-	-	-	-
S <sub>5</sub>	-	-	S	-	X	-	-	-	S	-	-
H <sub>6</sub>	S	-	-	-	-	X	-	-	-	-	-
H <sub>7</sub>	-	-	S	-	-	-	X	-	-	-	-
H <sub>8</sub>	S	-	-	-	-	-	-	X	-	-	-
H <sub>9</sub>	-	-	-	-	S	-	-	-	X	-	-
H <sub>10</sub>	S	-	-	-	-	-	-	-	-	X	-
H <sub>11</sub>	-	-	S	-	-	-	-	-	-	-	X

Fragment codes capture predefined groups of atoms and bonds. This is the representation used in both the interactive and automatic design procedures. One major limitation typically expressed about fragment codes is that they do not provide information about the components of the group — the atoms and bonds. I defined groups to be objects giving access to this information.

#### 6.4.1 Groups

Groups are objects which represent their molecular structure by a list of atoms and bonds. Atoms and bonds are themselves objects. The relevant instance variables for groups, atoms, and bonds are shown in Table 6.1.

---

Table 6.1: Molecule Representation Objects

---

Object	Instance Variables
group	atoms-list, bonds-list, linear-name
md-atom	atom-string, bond-list
bond	connected-atom-1, connected-atom-2

---

Bond objects are the basis of my representation. Each bond is connected to two and only two atoms. These connected atom objects are stored in the `connected-atom-1` and `connected-atom-2` instance variables. The `atom-string` instance variable of the `md-atom` object stores a string denoting the type of atom. The string corresponds to the chemical symbols used for atoms. Table 6.2 shows all the atoms currently known to the system. The `bond-list` of an `md-atom` object stores a list of all connected bonds. This bond list is used to traverse a molecular structure finding connected atoms. Groups keep lists of atoms and bonds.

Some of the bonds in a group are not connected to atoms. To facilitate manipulation I created a “free” atom to which these bonds connect. This is to ensure that all bonds are connected to exactly two atoms. Free atoms are stored in the group’s `atoms-list` instance variable. They directly facilitate determining the global valence of a group. The number of free atoms corresponds to the global valence.

---

Table 6.2: All Atoms Currently Known to the System

---

Boron	Bromine	Carbon	Chlorine
Fluorine	Hydrogen	Iodine	Nitrogen
Oxygen	Phosphorous	Silicon	Sulfur

---

## 6.5 Model Entry Configuration Objects

The six major objects used in the Model Configuration are:

1. **group-contribution-input-object**
2. **estimation-technique-class**
3. **estimation-technique**
4. **equation-oriented-estimation-technique**
5. **group-contribution-estimation-technique**
6. **technique-input-object**

The definitions of these objects and their associated functions are in the files:

**molecular-design:group-contribution-section;objects.lisp**

and

**molecular-design:properties;estimation-techniques.lisp.**

The **technique-input-object** is used to store information input by the designer.

This information is used to create group contribution and equation oriented estimation techniques. Eight important instance variables store the required information:

1. **Type**
2. **Name**
3. **Estimated-Property**
4. **State-Variables**
5. **Required-Properties**
6. **Class**
7. **Estimation-Function**
8. **Interval-Estimation-Function**

The **type** instance variable stores whether the estimation technique being developed is group contribution or equation oriented. The estimation technique is referred by **name** and estimates the property stored in the **estimated-property** instance variable. The **class** instance variable stores the **estimation-technique-class** to which the estimation technique belongs. The class is used for categorizing estimation techniques so they can be displayed orderly. The **estimation-function** and **interval-estimation-function** contain the LISP code used to calculate the physical property estimate. Equation oriented estimation techniques can also require state variables such as temperature and pressure and require additional physical properties to estimate the estimated property. These values are stored in the **state-variables** and **required-properties** instance variables.

Group contribution estimation techniques require a set of groups and values for their contributions. The **group-contribution-input-object** stores and displays the input values for group contributions. The **Choose Group** and **Choose Groups** commands prompt the designer for a group and its contribution. These are stored in two instance variables of the **group-contribution-input-object**:

1. **group-object**
2. **value**

**Group-contribution-input-objects** are displayed in the Chosen Groups Pane. The group is displayed with the value for the contribution to its lower right.

The **estimation-technique-class** is used to organize the presentation of estimation techniques. The single important instance variable, **label-string**, stores the name used for display.

**Group-contribution-estimation-technique** and **equation-oriented-estimation-technique** objects are both built on the **estimation-technique** object. This flavor contains five major instance variables:

1. **estimated-property**
2. **required-properties**
3. **estimation-function**
4. **interval-estimation-function**
5. **class**

These instance variables store the same information as their analogs in the **technique-input-object**.

**Equation-oriented-estimation-technique** objects contain one additional instance variable: **state-variables**. This stores a list of the form (temperature), (pressure), or (temperature pressure). **Group-contribution-estimation-technique** objects contain two additional instance variables:

1. **group-data-vector**
2. **intercept**

The **group-data-vector** contains an association list of groups and contributions. The **intercept** instance variable contains the value for the intercept in the group contribution model. This is used by in the interactive design to establish the starting point.

## 6.6 Editing Configuration Commands

The following commands are listed in the Command Menu of the editing Configuration.

The definitions of these commands are in the file:

**molecular-design:group-contribution-section;commands.lisp.**

**Align to Grid:** Repositions each of the atoms in the Group Edit Pane so that its x-center and y-center coordinates are an integer multiple of the grid size.

**Clear Editing Pane:** Removes all objects from the Group Edit Pane.

**Connect Atoms:** Forms a bond between two atoms. The type of bond is that currently designated in the Bond Type Pane. The connection begins by choosing the first atom. After clicking on this atom a bond is tracked from the atom to the mouse. Clicking on another atom causes the two atoms to be connected. Clicking elsewhere causes the bond to be considered a free bond.

Middle clicking on the second choice repeats the connection procedure. In this manner a number of atoms can be quickly connected.

**Group Objects:** This command collects the selected bonds and atoms of the Group Edit Pane and forms them into a Molecular Structure object.

**Hide Atoms:** For many molecules it is desirable for appearance sake not to display all the atoms in the molecule. Benzene is perhaps the most common example in which the carbon and hydrogen atoms are not displayed. This command hides all the selected atoms. The atoms are still part of the group but are not displayed.

**Input Atoms:** This command prompts for a sequence of atom names. An md-atom object is created for each atom name entered. These atoms are added to the *Group Edit Pane*.

**Mass Deselect Atoms:** All atoms within a rectangular area specified by the mouse are deselected. See the command Mass Select Atoms.

**Mass Deselect Bonds:** All bonds within a rectangular area specified by the mouse are deselected. See the command Mass Select Atoms.

**Mass Select Atoms:** Some commands, e.g. Hide Atoms, operate only on selected atoms. This command lets the designer select all the atoms in an area of the screen using the mouse. After activation of this command the mouse cursor turns into the upper left hand corner of a rectangle. Clicking the mouse places this rectangle corner at the current mouse position. The mouse cursor then changes to a lower right hand corner. A “rubber-banding” box is displayed connecting the two corners. Clicking the mouse again fixed the second corner thus specifying the area enclosed by the rectangle. All atoms in this area are selected.

**Mass Select Bonds:** All bonds within a rectangular area specified by the mouse are selected. See the command Mass Select Atoms.

**Unhide Atoms:** Redisplays the atoms which were hidden using the Hide Atoms command.

## 6.7 Model Entry Configuration Commands

The following commands are listed in the Command Menu of the Model Entry Configuration. The definitions of these commands are in the file:

**molecular-design:group-contribution-section;commands.lisp.**

**Activate Technique:** Prompts the designer for a **technique-input-object**. The command collects all the information stored in the object and all the **group-contribution-input-objects** from the Chosen Groups Pane and forms an estimation technique. This technique is now available for use throughout the system.

**Choose Group:** This command is used to enter the contribution for a group. The system prompts the designer to choose a group from the All Groups Pane. The system next prompts for the contribution of this group. After entry the system displays the chosen group with its contribution in the Chosen Groups Pane.

**Choose Groups:** This command is analogous to the **Choose Group** command. However, the system enters a loop which reprompts the designer to enter another group and its contribution.

**Create Estimation Technique:** Creates a new **technique-object**, prompts for initial information, and adds the object to the Technique Description Pane. The information prompted for is:

1. Type
2. Name
3. Class
4. Estimated Property
5. State Variables
6. Required Properties

This information can be changed by using the **Edit Technique** command.

**Create Estimation Technique Class:** Prompts the designer for the name of the new class. This name is used to form a new `estimation-technique-class` object. The command then prompts if the new class is to be saved to file. The file in which classes should be stored is:

```
molecular-design:properties;technique-class-instances.lisp.
```

**Delete Estimation Model:** Removes the estimation model from a chosen technique object. The command first prompts for a technique object. The estimation model is removed from the chosen technique. The Technique Description Pane is then redisplayed showing the value has been removed.

**Delete Interval Model:** Removes the interval model from a chosen technique object. The command first prompts for a technique object. The interval model is removed from the chosen technique. The Technique Description Pane is then redisplayed showing the value has been removed.

**Edit Documentation:** `Technique-input-objects` contain documentation which is stored in any estimation techniques created. This command first prompts for a `technique-input-object`. The current documentation for the chosen object is placed into the Entry Pane editor. Pressing the `<END>` key at the completion of editing causes the edited string to be extracted from the editor and stored back into the `technique-input-object`.

**Edit Estimation Model:** This command activates the editor used to enter the technique's estimation model. The Entry Pane is a ZMACS editor providing complete editing facilities. Typing <END> exits the editor and stores the entered model in the **technique-object**'s Estimation Model slot.

The **group-sum** special variable is used in an estimation model to represent the sum of the group contributions for a particular molecule. The system automatically creates the code necessary to assign the correct value to this variable. Documentation on the **group-sum** special variable is presented in a window resource exposed over both the All Groups Pane and Chosen Groups Pane while the editor is active.

**Edit Input Object:** Reprompts the designer with the menu used to originally create the **estimation-technique-object**. The menu prompts for five values:

1. Type.
2. Name.
3. Estimated Property.
4. State Variables.
5. Required Properties.

The command first prompts for a technique object. A menu is exposed which contains the current values for the chosen technique object available for editing. Once the designer has made his or her entries the object is updated and the Technique Display Pane is redisplayed to show the new values.

**Edit Interval Model:** This command activates the editor used to enter the technique's interval model. See the **Edit Estimation Model** command.

**Redisplay All Input Groups:** Because of group specification or the deletion of individual groups, the groups displayed in the All Groups Pane may not represent all the groups known to the system. This command removes all the groups currently in the All Groups Pane and adds back all the groups known to the system. The All Groups Panes is then redisplayed showing all these groups.

**Redisplay Techniques Pane:** Redisplays each of the seven panes of the Model Entry Configuration.

**Remove All Chosen Groups:** Removes all the groups whose contributions have been specified. The Chosen Groups Pane is cleared. Removing all groups is necessary when contributions are to be specified for a new group contribution estimation technique.

**Save Technique:** This command saves an estimation technique to file. The command prompts for a **technique-object**. The information stored in this object is used to form an estimation technique which is stored to file. The file in which estimation techniques should be stored is:

```
molecular-design:properties;technique-instances.lisp.
```

**Show All Groups:** Displays all groups known to the system in the All Groups Pane. Use this command to see any newly created groups.

**Specify Input Groups:** Most group contribution estimation techniques do not specify contributions for all the groups known to the system. For example, the group

contribution estimation techniques for the Polar and Hydrogen Solubility Parameters contain no contributions for cyclic groups. This command facilitates the selection of groups by allowing the designer to remove all cyclic groups from the All Groups Pane. The **Redisplay All Input Groups** command redisplays all the groups known to the system in the All Groups Pane.

**Verify Technique Functions:** This command simply checks if all the symbols in the entered models are known to the system. Models can include any LISP function, the special variables: `group-sum`; `meta-group-sum`; `check-interval-values`, and any physical properties known to the system.

The system first prompts for a `technique` object. The estimation technique's models are checked. If both models are verified the system displays a message stating this. If one or both models contain symbols which are not recognized by the system they are collected and presented to the designer in a warning message.

## 6.8 Section Discussion

The current interface for entering groups and models is somewhat cumbersome. The All Groups Pane of the Model Entry Configuration should be changed to a AVV format. This necessitates using linear names when prompting for groups. A standard notation for linear names needed.

Creating new groups is time consuming. Providing additional facilities for the positioning of atoms would reduce the time required.

### 6.8.1 Estimation Techniques

The current necessity of entering two estimation models is cumbersome. Algebraic rearrangement of the equations used in estimation can significantly reduce excess interval width. Thus it is necessary to have an estimation model specifically oriented toward interval evaluation. However, interval arithmetic reduces to conventional arithmetic when the intervals are of zero width, i.e., real numbers. The interval oriented estimation models are perfectly satisfactory for use as estimation models. I thus recommend that the models used in both group contribution and equation oriented estimation techniques be replaced by a single interval oriented model.

The current collection of interval arithmetic functions should be rewritten. These functions currently change all arguments to intervals before performing any interval arithmetic manipulations. To ensure efficiency is not compromised when interval functions are called with real arguments I recommend the arguments be checked and if found to be real arguments then the default arithmetic functions be used. This checking of argument types for function dispatch should be easily handled by generic functions.

An alternative method to the use of interval oriented models is to use standard arithmetic models and use some of the algorithms for interval width reduction to tighten interval bounds. The one disadvantage this has is that the interval tightening is done at run time. Algebraic rearrangements are done before a run begins.

I recommend that research be done to investigate how equations can be rearranged to reduce excess interval width. Identification of monotonic terms within a function can provide a large reduction in excess width. Creating unified extensions of nonmonotonic

terms also leads to the reduction of excess width.

### 6.8.2 Molecule Representation

I believe that the representation I use for molecular structure is very good. It is simple in structure yet quite expressive. I also believe that chiral structures can be represented.

The position of the atoms in the `atoms-list` of a group can be used to capture this information.

Further research in representing complete molecules is needed. Investigating the ability of the representation to capture chirality should be a priority.

## 6.9 Example Usage

In this example we create a new property, create new groups, and enter a group contribution estimation technique. The Problem Formulation Section provides facilities for property creation. The Group Editing and Model Entry configurations of the Group Contribution Section provide facilities for the creation of groups and entry of estimation techniques.

### Changing Configurations

New properties are created in the Problem Formulation Section. If the current configuration is not the Problem Formulation Section the first task is to change configurations.

**Action 6.1** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 6.2** *Mouse left on the Problem Formulation Section.*

The system changes configuration to the Problem Formulation Section Constraints Configuration.

### **Creating a Property Class**

If the new property does not belong to an existing class then a new property-class object must first be created.

**Action 6.3** *Mouse left on the Create Property Class command.*

The system prompts for the name of the new property class.

**Enter the name of the new class:**

**Action 6.4** *Type Magnetic Properties. Press return.*

The system next queries if the designer wishes to save this new property class to a file.

**Do you wish to save this new class to file?**

**Action 6.5** *Type No. Press return.*

The Property List Pane is redisplayed showing the new property class.

## Creating Properties

Now that we created a new property class we create the new property.

**Action 6.6** *Mouse left on the Create Property command.*

The system prompts for the following information:

**Pretty Name:** *some value*

**Short Name:** *some value*

**Variable Dependencies:** *Temperature Pressure*

**Default Minimum:** *some value*

**Default Maximum:** *some value*

**Property Class:** *some value*

**ABORT** *aborts*, **END** *uses these values*

**Action 6.7** *Mouse left on the phrase after the Pretty Name: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.8** *Type Molar Diamagnetic Susceptibility. Press return.*

**Action 6.9** *Mouse left on the phrase after the Short Name: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.10** *Type X. Press return.*

Variable Dependencies denotes the state variables this property is dependent upon.

At present the possible state variables are Temperature and Pressure. Our molar diamagnetic susceptibility is not dependent upon either state variable.

**Action 6.11** *Mouse left on the phrase after the Default Minimum: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.12** *Type 0. Press return.*

**Action 6.13** *Mouse left on the phrase after the Default Maximum: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.14** *Type 150. Press return.*

**Action 6.15** *Mouse left on the phrase after the Property Class: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.16** *Mouse left on our newly created property class, Magnetic Properties, displayed in the Property List Pane.*

**Action 6.17** *Press the <END> key.*

The new property is created and added to the Property List Pane under the Magnetic Properties property class. The **Save Property To File** command can be used to save the new property to a file. This command prompts the designer for a file in which to save the property. The file which contains the system properties is:

```
molecular-design:properties;property-instances.
```

In this example we do not save the property.

## Creating New Groups

Different group contribution estimation techniques require different groups. The Editing Configuration of the Group Contribution Section provides facilities for creating new groups. The group contribution estimation technique for molar diamagnetic susceptibility does not require any groups in addition to those already in the system. However, I go through this example to detail the operation of the configuration.

**Action 6.18** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 6.19** *Mouse left on the Group Contribution Section Editing Configuration.*

The system changes configuration to the Editing Configuration of the Group Contribution Section.

The Group Edit Pane is the canvas on which we connect and arrange atoms and bonds. We construct the group  $-(C_6H_4)CH_2CH_2CH_2-$ . We begin by entering the necessary atoms.

**Action 6.20** *Mouse left on the Input Atoms command.*

The system prompts for a sequence of atoms.

**Please enter one or more Atoms:**

**Action 6.21** *Type the following: Carbon Carbon Carbon Carbon Carbon Carbon Carbon Carbon Hydrogen Hydrogen. Press return.*

The system creates an `md-atom` object for each entry and adds it to the Group Edit Pane. Figure 6.4 shows the Group Edit Pane with the input atoms.

The atoms are now arranged into an aesthetically pleasing yet chemically informative structure. Atom movement is accomplished by the following action.

**Action 6.22** *Mouse h-left on any atom.*

The mouse cursor “picks up” the atom object underneath it. The location of this atom object is changed by moving the mouse.

**Action 6.23** *Move the mouse to another location on the Group Edit Pane.*

The mouse “puts down” the object at the current location.

**Action 6.24** *Mouse left.*

In this manner rearrange the atoms into a structure similar to that shown in Figure 6.5.

To facilitate the alignment of atoms and bonds the Group Edit Pane maintains a grid. The current size of the grid is displayed in the upper left corner of the pane. Mousing left on the **Align to Grid** command repositions each of the atoms and bonds so its center is at the nearest grid point.

Atoms are connected using the **Connect Atoms** command. The type of bond used in the connection is specified by the Bond Type pane. The default, Single, is satisfactory.

**Action 6.25** *Mouse left on the Connect Atoms command.*

The command prompts for the first atom.

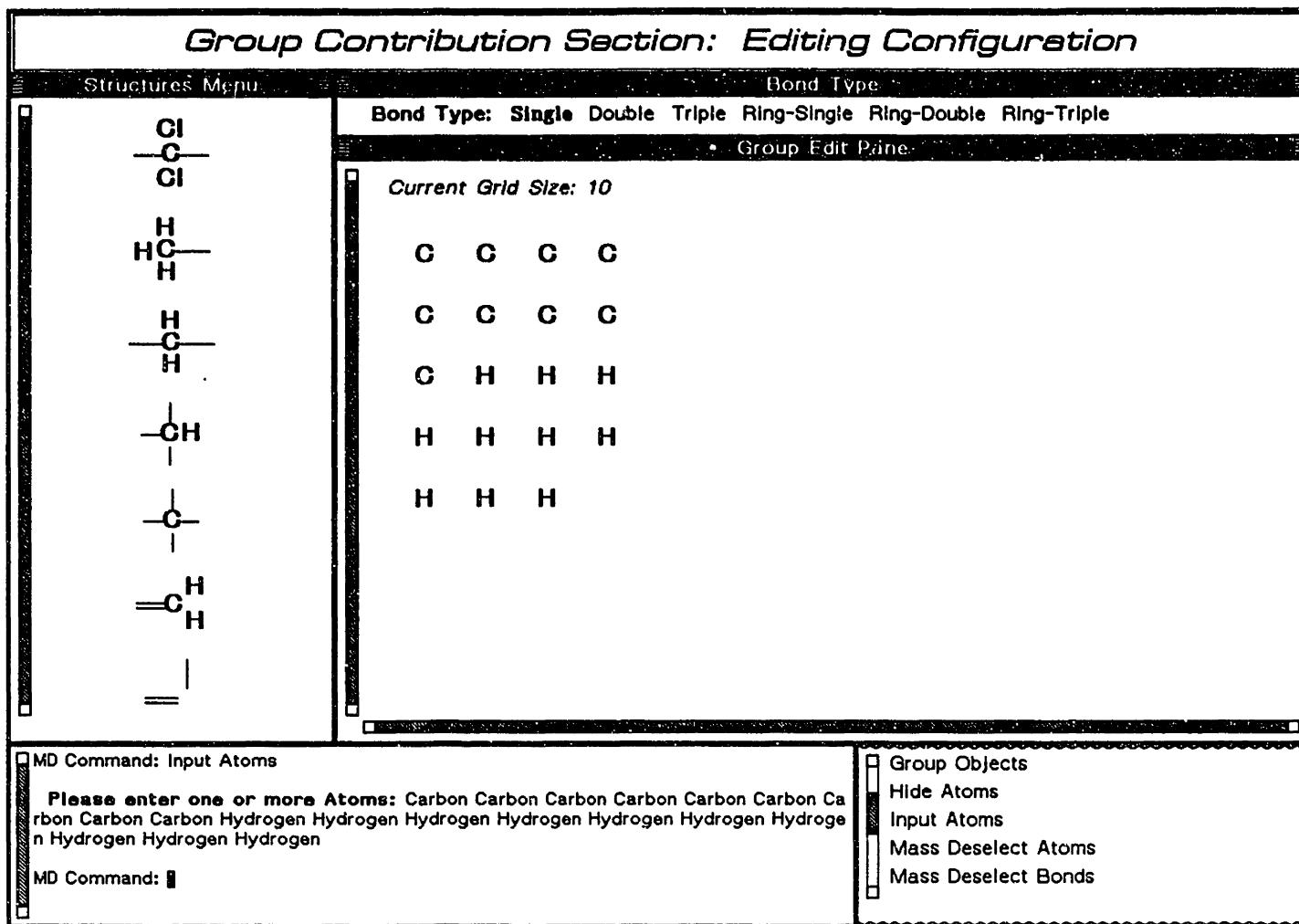


Figure 6.4: Input Atoms for the Construction of a New Group

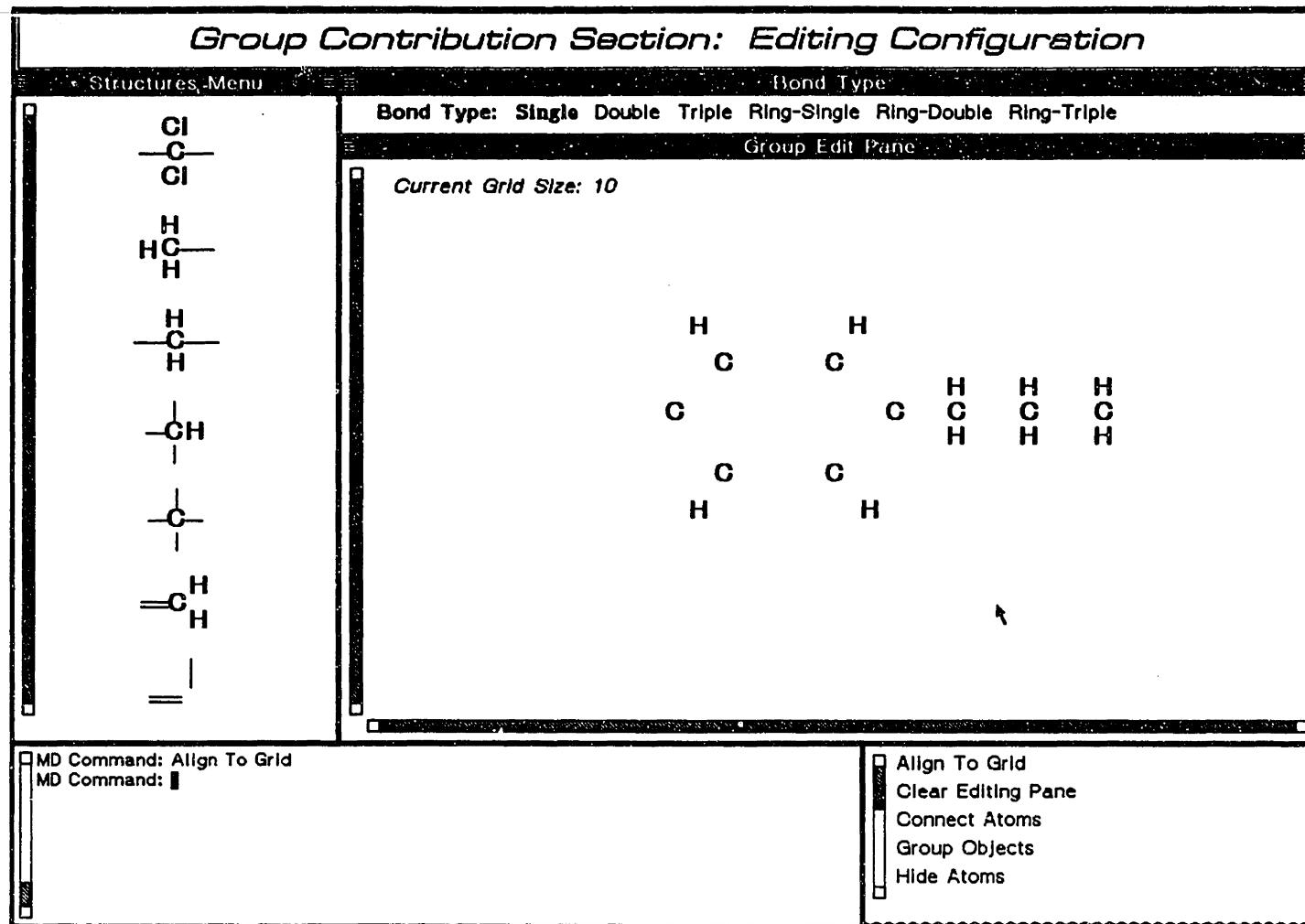


Figure 6.5: Input Atoms Rearranged into the Shape of the New Group

**Click on an Atom:**

**Action 6.26** *Mouse left on the carbon atom in the lower left corner.*

The system creates a single bond which has one end attached to the center of the atom and the other end attached to the mouse. The system prompts for the atom to connect.

**Click on another Atom or a Location:**

**Action 6.27** *Mouse left on the hydrogen atom in the lower left corner.*

The system connects the carbon and hydrogen atoms with a single bond.

Mousing middle on the second atom, instead of mousing left, repeats the connection procedure. In this manner the designer can continue to connect atoms without reinvoking the **Connect Atoms** command.

Connect the atoms of our Group Edit Pane so they appear as displayed in Figure 6.6. It is important to connect the carbons and hydrogens on the alkane branch. These bonds are too short to be drawn but still must be entered for the system to reason about the group properly.

The bonds used in a ring are denoted differently than those used in acyclic structures. We connect the carbon atoms in our ring first by ring-single bonds.

**Action 6.28** *Mouse left on Ring-Single displayed in the Bond Type Pane.*

The system denotes that Ring-Single is now the default bond type by redisplaying the list of bond types with Ring-Single typed in bold. Connect the carbons atoms in our ring as before. The resulting structure should look like Figure 6.7.

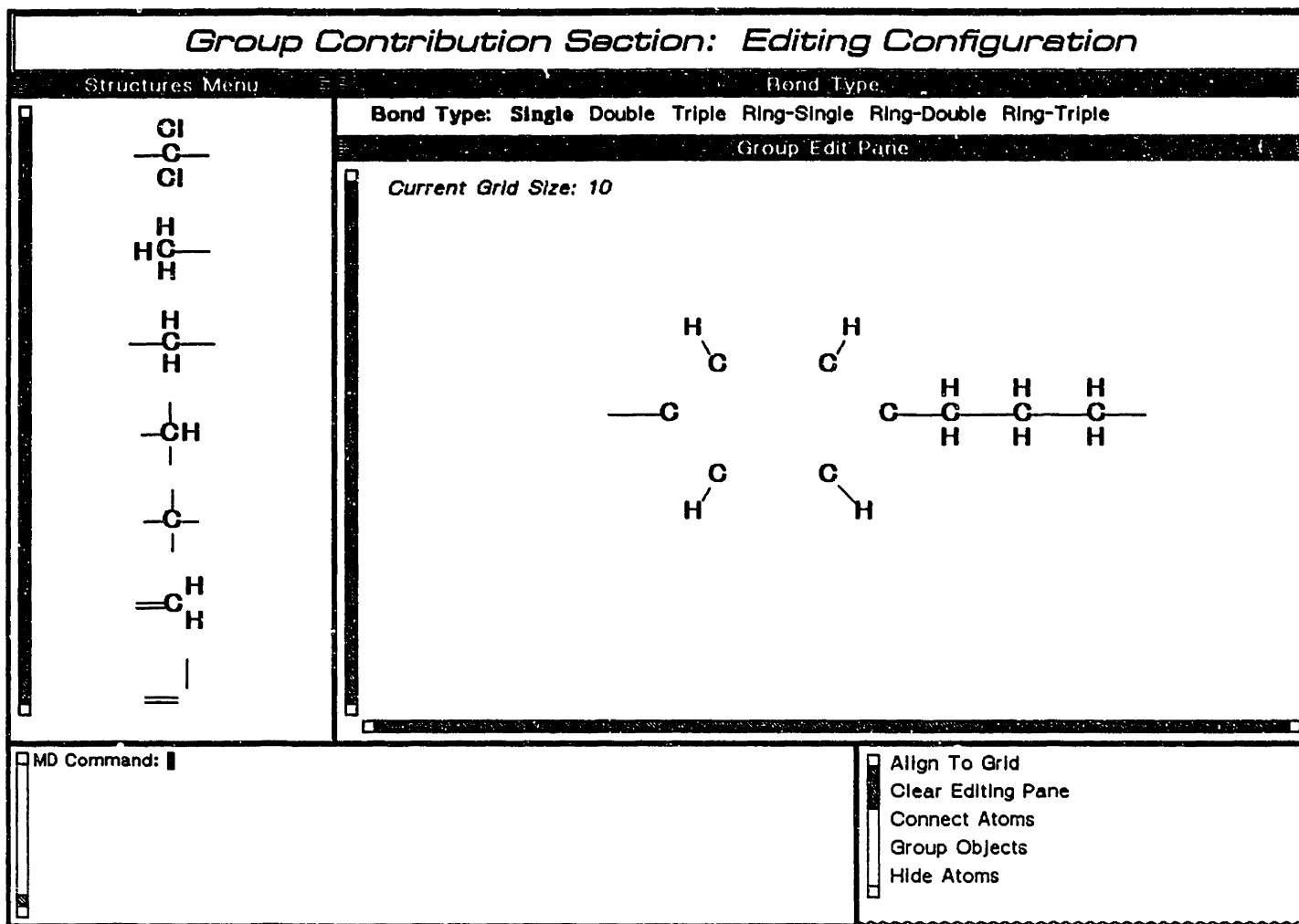


Figure 6.6: Single Bond Connections

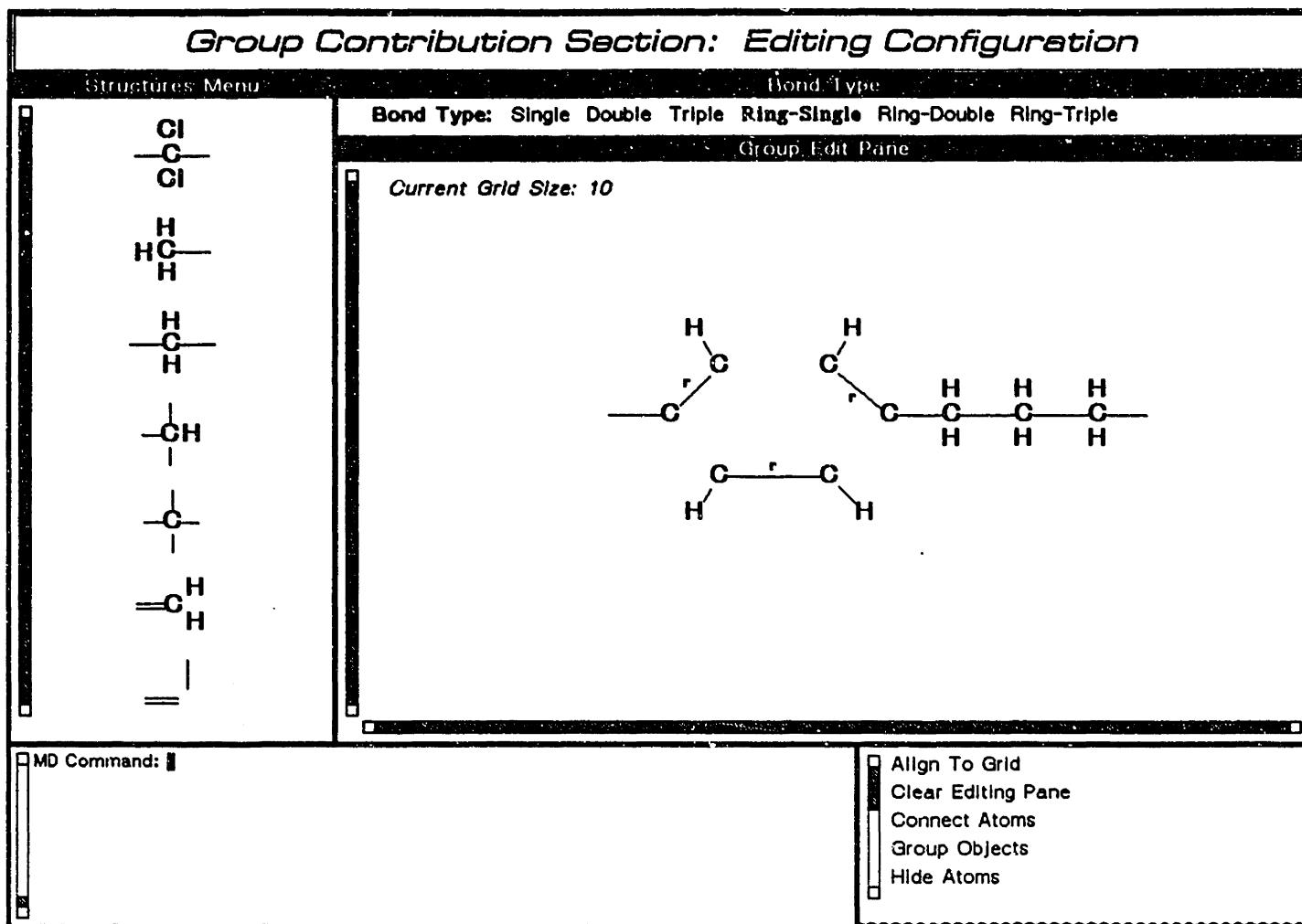


Figure 6.7: Ring Single Bond Connections

Repeating the procedure one more time to connect ring-double bonds gives us the structure shown in Figure 6.8.

For many molecular structures the detail of all the atoms, especially hydrogen atoms, leads to a cluttered presentation. The Editing Configuration provides facilities for “hiding” atoms. Although these atoms are still part of the molecular structure they are not drawn when the structure is presented.

**Action 6.29** *Mouse left on the Mass Select Atoms command.*

The command requires the designer to enclose a set of atoms using the “rubber-banding box” procedure. Those atoms enclosed by the box are selected.

**Action 6.30** *Enclose all the aromatic carbons and their attached hydrogens by the “rubber-banding box”.*

**Action 6.31** *Mouse left on the Hide Atoms command.*

The selected atoms are now drawn with a dashed box around them.

We now collect all the atoms into a new group.

**Action 6.32** *Mouse left on the Mass Select Atoms command.*

**Action 6.33** *Enclose all the atoms within the “rubber-banding box”.*

**Action 6.34** *Mouse left on the Mass Select Bonds command.*

**Action 6.35** *Enclose all the bonds within the “rubber-banding box”.*

**Action 6.36** *Mouse left on the Group Objects command.*

### Group Contribution Section: Editing Configuration

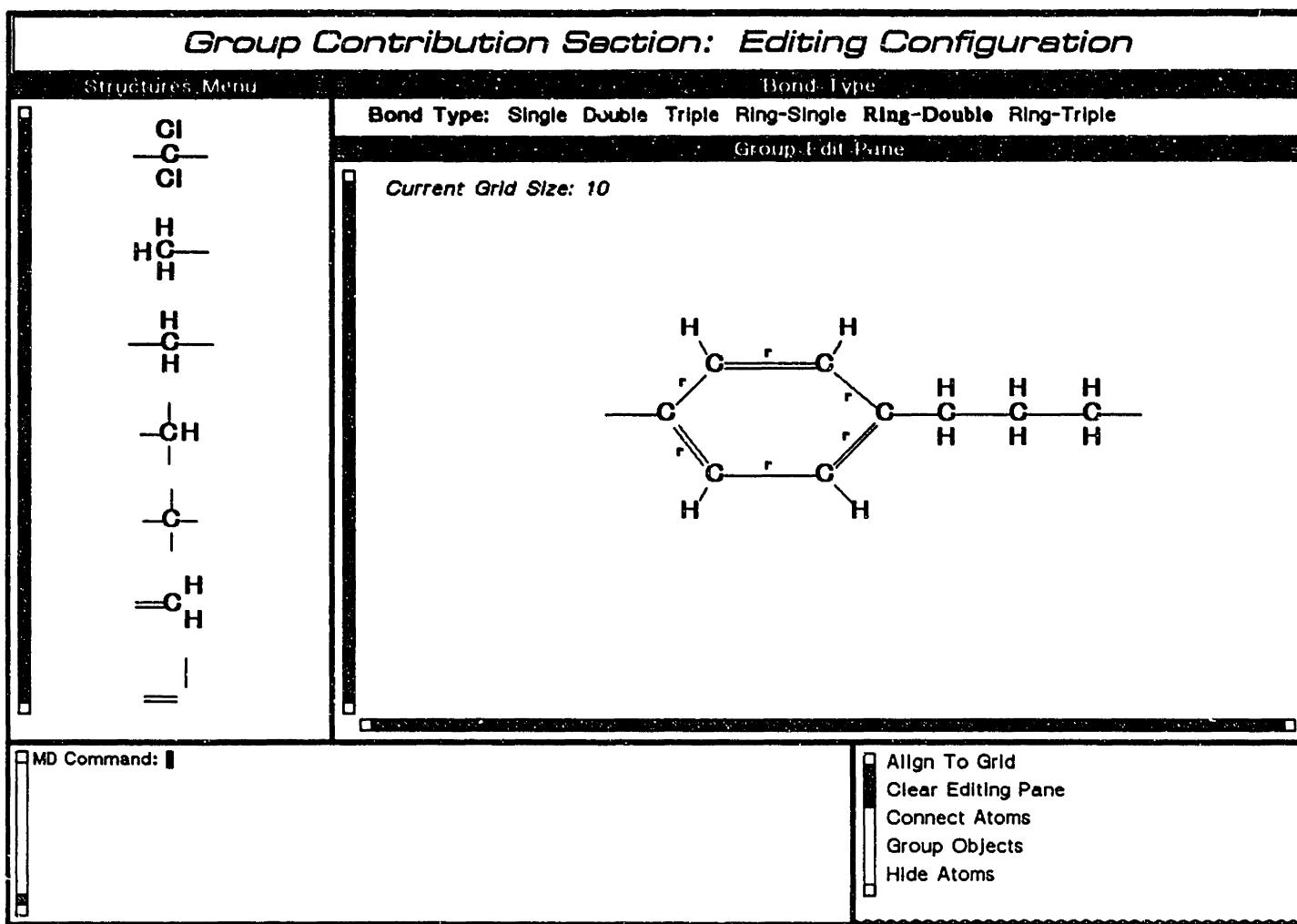


Figure 6.8: Completely Connected Input Atoms

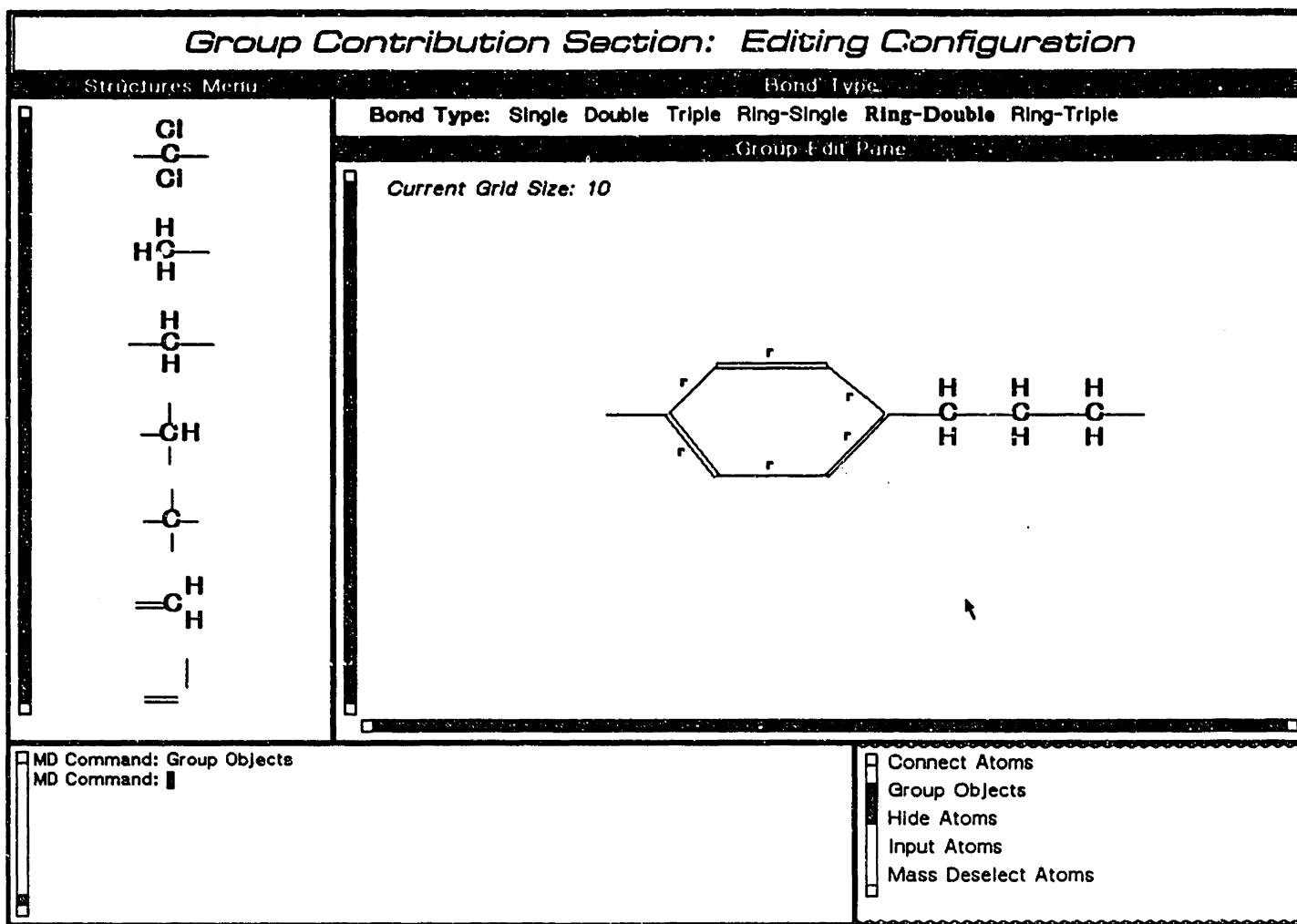


Figure 6.9: Input Atoms Collected into a New Group

All the atoms and bonds are collected into a new group object. This object is similar to that displayed in Figure 6.9.

The **Save Group** command activates this group and optionally saves it to a file.

**Action 6.37** *Mouse left on the Save Group command.*

The system prompts for a molecular structure:

**Click on a group:**

**Action 6.38** *Mouse left on the group displayed in the Group Edit Pane.*

The system prompts for a linear name for this group:

**Enter the Linear Name for this Group:**

**Action 6.39** *Type -(C6H4)CH2CH2CH2-. Press return.*

The system next prompts if you wish to save this group to a file.

**Do you wish this group saved to a file?**

**Action 6.40** *Type No. Press return.*

The Structures Menu is redisplayed with our new group at the top.

### **Creating an Estimation Technique Class**

We create a new estimation technique class for the estimation techniques we will develop. We first move to the Model Entry Configuration.

**Action 6.41** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 6.42** *Mouse left on the Group Contribution Section Model Entry Configuration.*

The system changes configuration to the Group Contribution Section's Model Entry Configuration.

**Action 6.43** *Mouse left on the Create Estimation Technique Class command.*

The system prompts for the name of the new estimation technique:

**Enter the name of the new class:**

**Action 6.44** *Type Magnetic Techniques into the interaction pane. Press return.*

The system prompts if you wish to save the new class to a file:

**Do you wish to save the new class to file?**

**Action 6.45** *Type No. Press return.*

The new estimation technique class is now available for input during the development of new estimation techniques.

## Entering an Estimation Technique

Estimation techniques require models and group contributions. The Model Entry Configuration of the Group Contribution Section provides the interface for entering new techniques. The Model Entry Configuration uses a **technique-input-object** to store the information needed to construct a new estimation technique.

**Action 6.46** *Mouse left on the Create Estimation Technique command.*

The system exposes a menu prompting for the following information:

**Type:** Group Contribution    Equation Oriented

**Name:** *some value*

**Class:** *some value*

**Estimated Property:** *some value*

**State Variables:** *some value*

**Required Properties:** *some value*

**Action 6.47** *Mouse left on Group Contribution following the Type: prompt.*

**Action 6.48** *Mouse left on the phrase following the Name: prompt. The phrase is replaced by a blinking cursor.*

**Action 6.49** *Type vanKrevelen X. Press return.*

**Action 6.50** *Mouse left on the phrase following the Class prompt. The phrase is replaced by a blinking cursor.*

**Action 6.51** *Mouse right on an empty space within the menu.*

The system exposes a menu containing known estimation technique classes.

**Action 6.52** *Mouse left on Magnetic Techniques.*

**Action 6.53** *Mouse left on the phrase following the Estimated Property prompt.*

*The phrase is replaced by a blinking cursor.*

**Action 6.54** *Mouse right on an empty space within the menu.*

The system exposes a menu containing known properties.

**Action 6.55** *Mouse left on Molar Diamagnetic Susceptibility.*

Our group contribution technique requires no additional properties and is not dependent upon temperature or pressure.

**Action 6.56** *Press the <end> key.*

The system creates a technique-input-object displaying it in the Technique Description Pane.

Since our new estimation technique is a group contribution technique we must enter the group contributions.

**Action 6.57** *Mouse left on the Choose Groups command.*

The system begins a loop prompting for groups and their contributions.

**Click on a Group:**

**Action 6.58** *Mouse left on the  $-CH_3$  displayed in the All Groups Pane.*

**Enter the Contribution Value:**

**Action 6.59** *Type 14.5. Press return.*

The system queries if you wish to continue, end, or redisplay the Chosen Groups Pane.

**Continue, End, or Redisplay:**

The choice is made by entering C, E, or R.

**Action 6.60** *Type E. Press return.*

Figure 6.10 shows the screen display after the entry of the group contribution. Repeat the above procedure for the group contributions shown in Table 6.3.

The second part of our group contribution estimation technique is the model.

**Action 6.61** *Mouse left on the Edit Estimation Model command.*

The system prompts for a technique-input-object.

**Enter a Technique Input Object:**

**Action 6.62** *Mouse left on our new technique-input-object displayed in the Technique Description Pane.*

The Entry Pane is activated indicated by the blinking cursor. The Entry Pane is a ZMACS-like editor. Editing is terminated by pressing the <END> key. Pressing the <ABORT> key aborts the editing.

The system displays some documentation detailing the keywords which can be used within models. Figure 6.11 shows a screen displaying the documentation.

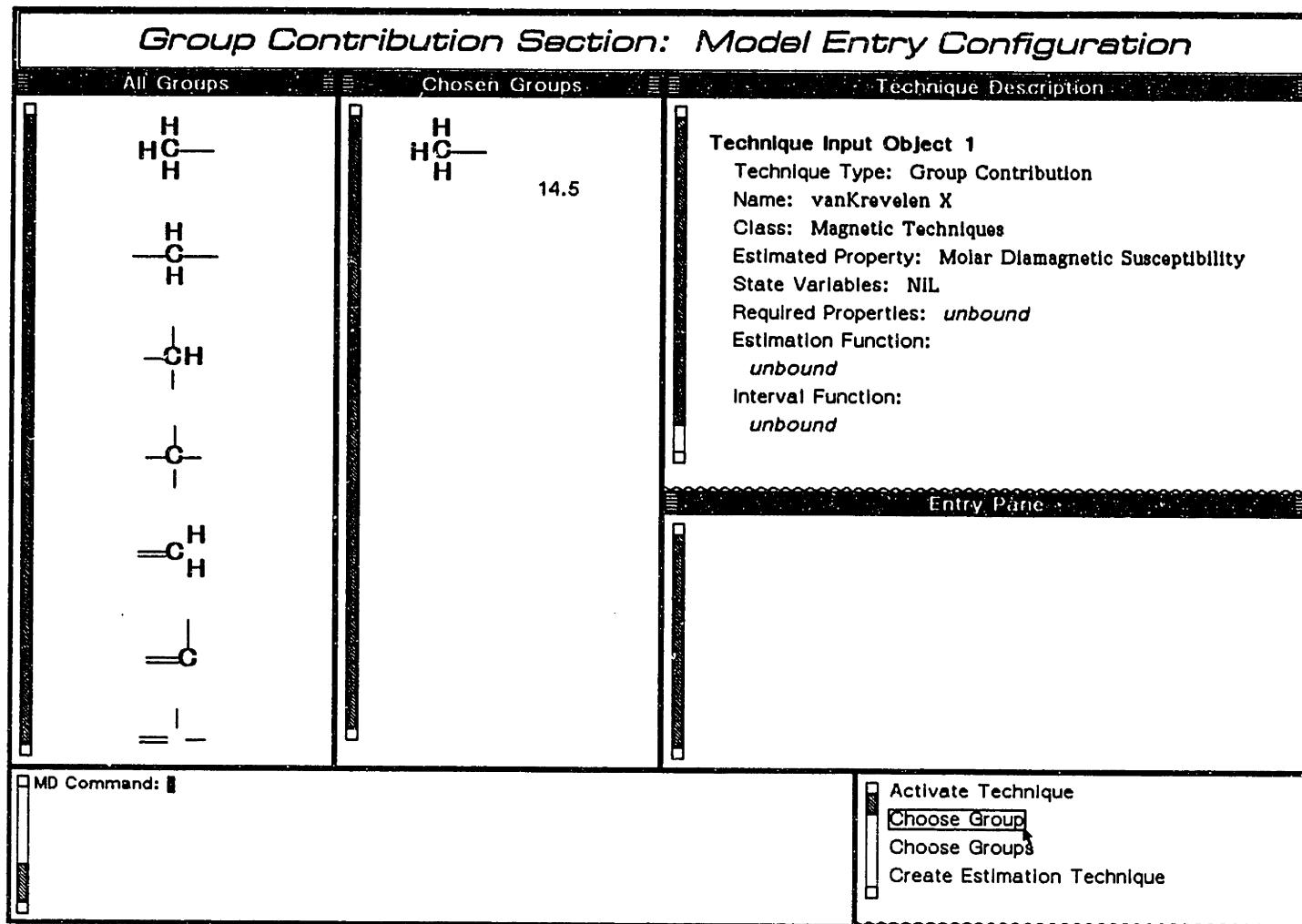


Figure 6.10: Estimation Technique Development for Diamagnetic Susceptibility

Table 6.3: Molar Magnetic Susceptibility Contributions

Groups	Contribution
<b>Acyclic Increments</b>	
$-\text{CH}_3$	14.5
$-\text{CH}_2-$	11.35
$>\text{CH}-$	9.0
$>\text{C}<$	7.0
$=\text{CH}_2$	9.0
$=\text{CH}-$	6.6
$=\text{C}<$	4.5
$\equiv\text{CH}$	9.0
$\equiv\text{C}-$	7.0
<b>Halogen Increments</b>	
$-\text{F}$	6.6
$-\text{Cl}$	18.5
$-\text{Br}$	27.5
$-\text{I}$	43.0
<b>Acyclic Oxygen Increments</b>	
$-\text{OH}$	7.5
$-\text{O}-$	5.0
$>\text{CO}$	6.5
$-\text{CHO}$	8.4
$-\text{COOH}$	19.0
$-\text{COO}-$	14.0
<b>Acyclic Nitrogen Increments</b>	
$-\text{NH}_2$	12.0
$>\text{NH}$	9.0
$>\text{N}-$	6.0
$-\text{CN}$	11.0
$-\text{NO}_2$	8.0
<b>Acyclic Sulfur Increments</b>	
$-\text{SH}$	18.0
$-\text{S}-$	16.0

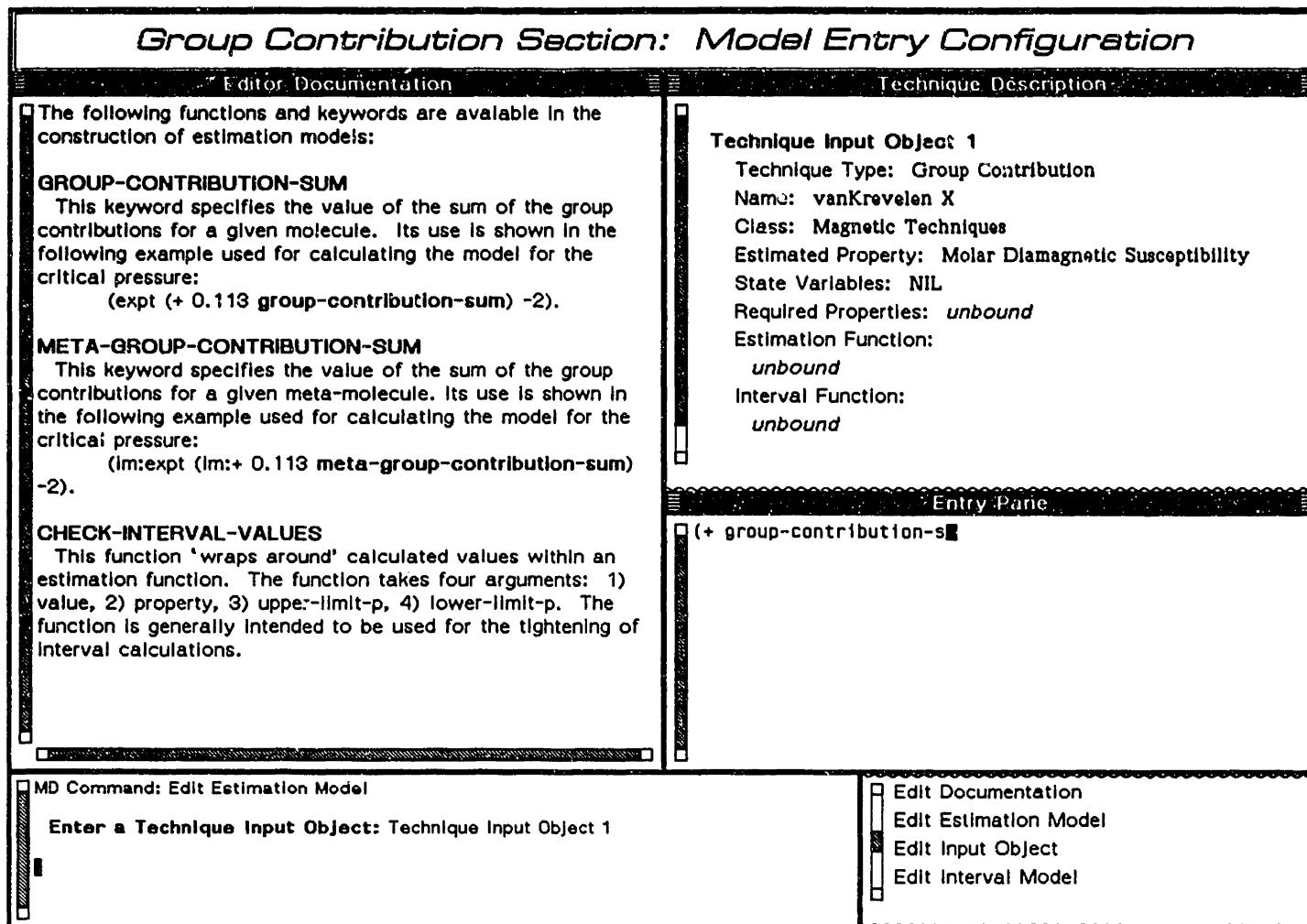


Figure 6.11: Editing an Estimation Model

**Action 6.63** *Type (+ group-contribution-sum 0.0) into the Entry Pane.*

**Action 6.64** *Press the <END> key.*

The new model is stored in our **technique-input-object**.

An interval estimation model is also entered.

**Action 6.65** *Mouse left on the Edit Interval Model command.*

The system prompts for a **technique-input-object**.

**Enter a Technique Input Object:**

**Action 6.66** *Mouse left on our technique-input-object displayed in the Technique Description Pane.*

**Action 6.67** *Type (im:+ meta-group-contribution-sum 0.0) into the Entry Pane.*

**Action 6.68** *Press the <END> key.*

The new interval model is stored in our **technique-input-object**.

## **Activating the Technique**

The final step in entering a new estimation technique is to activate the technique.

**Action 6.69** *Mouse left on the Activate Technique command.*

The system prompts for a **technique-input-object**.

**Enter a Technique Input Object:**

**Action 6.70** *Mouse left on our technique input object.*

The system creates the new estimation technique responding with:

X → () – vanKrevelen X – Group Contribution Technique has been activated.

# Chapter 7

## Target Transformation Section

Once physical property constraints are entered in the Problem Formulation Section, it is necessary to instruct the system how to estimate the properties contained in these constraints. This involves the creation of estimation procedures for each of the physical properties used in the design constraints. The Target Transformation Section provides facilities for collecting estimation techniques into estimation procedures.

### 7.1 Section Layout

The screen layout of the Target Transformation Section is shown in Figure 7.1. The screen real estate is used by five panes:

**Target Transformation Section Title Pane:** Displays the title of the Target Transformation Section.

## Target Transformation Section

Transformed Constraints Estimation Techniques Selection

**Acentric Factor Techniques**  
 $P_c \rightarrow (F_1 F_2 F_3)$  - Joback  $P_c$  Factor - Equation Oriented Technique  
 $P_c^* \rightarrow ()$  - Joback  $P_c^*$  - Group Contribution Technique  
 $P_c \rightarrow (P_c^*)$  - Joback  $P_c$  Modification - Equation Oriented Technique  
 $V_c \rightarrow ()$  - Joback  $V_c$  - Group Contribution Technique

**Critical Properties Techniques**  
 $P_c \rightarrow (F_1 F_2 F_3)$  - Joback  $P_c$  Factor - Equation Oriented Technique  
 $P_c^* \rightarrow ()$  - Joback  $P_c^*$  - Group Contribution Technique  
 $P_c \rightarrow (P_c^*)$  - Joback  $P_c$  Modification - Equation Oriented Technique  
 $V_c \rightarrow ()$  - Joback  $V_c$  - Group Contribution Technique

**Enthalpy Techniques**  
 $\Delta H_{298} \rightarrow ()$  - Joback  $\Delta H_{298}$  - Group Contribution Technique  
 $\Delta H_{v,b} \rightarrow (F_1 F_2 F_3)$  - Joback  $\Delta H_{v,b}$  Factor - Equation Oriented Technique  
 $\Delta H_{v,b} \rightarrow ()$  - Joback  $\Delta H_{v,b}$  - Group Contribution Technique  
 $\Delta H_v \rightarrow (\Delta H_{v,b} T_c T_{br})$  - Fish Lielmezs - Equation Oriented Technique  
 $\Delta H_v \rightarrow (\Delta H_{v,b} T_c T_b)$  - Watson Relation  $T_c$  Biased - Equation Oriented  
 $\Delta H_v \rightarrow (\Delta H_{v,b} T_{br} T_b)$  - Watson Relation  $T_{br}$  Biased - Equation Oriented

**Factor Techniques**  
 $F_2 \rightarrow ()$  - F2 Assumption - Equation Oriented Technique  
 $F_3 \rightarrow ()$  - Joback F3 - Group Contribution Technique

MD Command:

Figure 7.1: Target Transformation Section Screen

**Transformed Constraints Pane:** Displays transformed constraints. As estimation techniques are chosen by the designer they are added to the selected transformed constraints in this pane.

**Estimation Techniques Selection:** Displays all the physical property estimation techniques known to the system. The estimation techniques are arranged by physical property class.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Target Transformation Section Commands Menu:** The command menu containing commands relevant to the Target Transformation Section.

## 7.2 Target Transformation Objects

The five major objects of the target transformation section are:

1. **transformed-constraint**
2. **estimation-technique-class**
3. **equation-oriented-estimation-technique**
4. **group-contribution-estimation-technique**

**Transformed-constraints** maintain information about the constraint value and the currently allocated estimation techniques. Both the **equation-oriented-estimation-technique** object and **group-contribution-estimation-technique** object are built from the **estimation-technique** object. From this structuring they inherit

a number of instance variables which store information about the estimated property, class, and estimation function. The **equation-oriented-estimation-technique** flavor adds a **state-variables** instance variable. **State-variables** keeps a list of state variables on which the estimated property is dependent. The system recognizes **temperature** and **pressure** as state variables. Only **temperature** is used in the current estimation techniques.

The **group-contribution-estimation-technique** flavor adds two instance variables:

1. **group-data-vector**
2. **intercept**

The **group-data-vector** is an association list of groups and their contributions. The **intercept** is the constant term in a linear group contribution model.

Table 7.1 shows all the estimation techniques currently known to the system arranged by **estimation-technique-class**. The name of an estimation technique appearing in the Estimation Techniques Selection Pane is of the form:

*Mapping - Technique Name - Technique Type*

The mapping denotes the physical property being estimated and the required physical properties. The mapping for the Watson Relation  $T_c$  Biased is:

$$\Delta H_v \rightarrow (\Delta H_{vb} \ T_c \ T_b) \quad (7.1)$$

means that the estimation “maps” a constraint on  $\Delta H_v$  into a constraint on  $\Delta H_{vb}$ ,  $T_c$ , and  $T_b$ . The techniques’ names are shown in Table 7.1. The technique type is either **Equation Oriented Technique** or **Group Contribution Technique**.

---

Table 7.1: All Estimation Techniques

---

**Acentric Factor Techniques**  
Lee Kesler Acentric Factor

**Critical Properties Techniques**  
Joback  $P_c$  Modification  
Joback  $P_c^*$   
Joback  $P_c$  Factor  
Joback  $V_c$

**Enthalpy Techniques**  
Joback  $\Delta H_{vb}$  Factor  
Joback  $\Delta H_{vb}$   
Fish Lielmezs  
Watson Relation  $T_c$  Biased  
Watson Relation  $T_{br}$  Biased  
Joback  $\Delta H_{f,298}^o$

**Factor Techniques**  
Joback  $F_1$   
Joback  $F_2$   
Joback  $F_3$   
 $F_2$  Assumption

**Temperature Techniques**  
Joback  $T_c$  Factor  
 $T_c$  Definition  
 $T_{br}$  Definition  
Joback  $T_{br}$  Modification  
Joback  $T_{br}^*$   
Joback  $T_m$   
Joback  $T_b$  Factor  
Joback  $T_b$

---

---

Table 7.1 Continued: All Estimation Techniques

---

**Heat Capacities Techniques**

$C_{pv}$  Cubic in Temperature  
van Krevelen  $C_{ps}$   
Joback  $C_{pv,a}$   
Joback  $C_{pv,b}$   
Joback  $C_{pv,c}$   
Joback  $C_{pv,d}$   
Joback  $C_{pv,298}$  Factor  
Rowlinson  $C_{pL}$

**Polymer Techniques**

van Krevelen Rao Function  
van Krevelen  $P_{LL}$   
Salame Permachor  
van Krevelen  $T_g$

**Vapor Pressure Techniques**

Riedel Plank Miller  $T_c$  Biased  
Riedel Plank Miller  $T_{br}$  Biased

**Volume Techniques**

van Krevelen  $V_S$

**Unspecified Techniques**

Joback Polar Solubility Parameter  
Joback Hydrogen Bonding Solubility Parameter

---

$T_b$ ,  $T_c$ , and  $T_{br}$  are related so that any one is equal to the quotient of the other two. Equation oriented estimation techniques whose names contain “bias” have had either  $T_c$  or  $T_{br}$  substituted. This was done to facilitate algebraic rearrangement of the estimation model. Such rearrangement is often useful in reducing the excess width in interval evaluation.

### 7.3 Target Transformation Commands

The following commands are listed in the Command Menu of the Target Transformation Section. The definitions of these commands are in the file:

```
molecular-design:target-transformation;commands.lisp.
```

**Apply Selected Techniques:** For each of the selected transformed constraints this command determines which of the selected estimation techniques would be applicable in forming the required estimation procedures. Those techniques which are found applicable are stored in the constraint object. Properties for which estimation techniques are chosen are removed from the **non-fundamental-properties** instance variable of the constraint object. The required properties of chosen equation oriented estimation techniques are added to the **non-fundamental-properties** instance variable. Properties for which group contribution estimation techniques were chosen are added to the **fundamental-properties** instance variable. All the selected constraints are applied one at a time. The Transformed Constraints Pane is redisplayed showing the updated **transformed-constraint** objects after all techniques have been applied to

all constraints.

**Check Dimensionality:** Checks the number of fundamental physical properties each selected constraint in the Transformed Constraint Pane has. If the dimension is greater than 3, this command presents the constraint's label to the designer. A constraint whose dimension is greater than 3 can not be processed by the interactive design method.

**Check Fundamentality:** Each of the selected constraints of the Transformed Constraints Pane is checked for the absence of non-fundamental properties. A transformed constraint can not be further processed until all properties have been reduced to fundamental properties. The transformed constraints which do not satisfy the fundamentality criterion are reported to the designer.

**Deselect all Techniques:** Deselects all the selected techniques displayed in the Estimation Techniques Selection Pane.

**Deselect all Transformed Constraints:** Deselects all the selected constraints displayed in the Transformed Constraints Pane.

**Display all Techniques:** Some commands reduce the techniques display in the Estimation Techniques Selection Pane. This command redisplays all the techniques known to the system.

**Display Applicable Techniques:** An applicable technique is defined as one whose estimated property is a non-fundamental property of one or more selected transformed

constraints. Each estimation technique is checked for applicability. When applicable it is collected into a list. Finally the displayed estimation techniques are removed from the Estimation Technique Selection Pane and the list of applicable techniques added.

**Display Applied Techniques:** Displays on a window resource exposed over the Estimation Techniques Selection Pane each selected transformed constraint with the estimation techniques which have been applied to it so far.

**Document Estimation Technique:** Prompts for an estimation technique and displays the technique's documentation in a window resource exposed over the Estimation Techniques Selection Pane. Figure 7.2 shows documentation for the Watson relation.

**Make Automatic:** This command prepares the transformed constraints for use in an automatic design. The selected constraints of the Transformed Constraints Pane are collected for preparation. Each constraint is checked for dependency only on fundamental properties. The valid constraints are then coerced into automatic design constraints and added to the Automatic Design Section: Meta-Molecules Configuration's Property Constraints Pane.

The system collects the group contribution estimation techniques used in all the estimation procedures. The groups used in each of these techniques are intersected to form a consistent group set. This group set is used to construct a meta-group node object and a leaf group. The meta-group node object is added to the Meta-Groups Configuration's Meta-Groups Display Pane. The leaf group is added to Meta-Groups Configuration's Leaf Groups Pane.

## Target Transformation Section

Transformed Constraints      Technique Documentation

**Transformed Constraint**  
Label: Cpl Small  
Value: (< (CPL 294.3) 32.2)  
Non-Fund: ( Cpl )  
Fund: ( None )

**Transformed Constraint**  
Label: Hv Large  
Value: (> (HV 272.05) 18.4)  
Non-Fund: ( ΔHv )  
Fund: ( None )

**Transformed Constraint**  
Label: Pvp High  
Value: (< (PVP 316.4) 14)  
Non-Fund: ( Pvp )  
Fund: ( None )

**Transformed Constraint**  
Label: Pvp Low  
Value: (> (PVP 272.05) 1.4)  
Non-Fund: ( Pvp )

The Watson relation is a widely used correlation between  $\Delta H_v$  and Temperature. It is typically used to estimate the enthalpy of vaporization at one temperature when it is known at another temperature. Here the known temperature is the normal boiling point and the corresponding value of the enthalpy of vaporization is the enthalpy of vaporization at the normal boiling point. The exponent of the equation has been found to vary from substance to substance. Reid et. al. [Reid77] has suggested that the constant value of 0.38 be chosen. Below  $T_b$ ,  $\Delta H_v$  increases more rapidly with decreasing temperature than predicted with a constant exponent.

Type any character to continue.

$\Delta H_{vb} \rightarrow ()$  - Joback Hvb - Group Contribution Technique  
 $\Delta H_v \rightarrow (\Delta H_{vb} T_c T_{br})$  - Fish Lielmezs - Equation Oriented Technique  
 $\Delta H_v \rightarrow (\Delta H_{vb} T_c T_b)$  - Watson Relation Tc Biased - Equation Oriented  
 $\Delta H_v \rightarrow (\Delta H_{vb} T_{br} T_b)$  - Watson Relation Tbr Biased - Equation Oriented

**Factor Techniques**  
 $F_2 \rightarrow ()$  - F2 Assumption - Equation Oriented Technique  
 $F_3 \rightarrow ()$  - Joback F3 - Group Contribution Technique

MD Command: Document Estimation Technique  
Enter an Estimation Technique:  $\Delta H_v \rightarrow (\Delta H_{vb} T_{br} T_b)$  - Watson Relation Tbr Biased

Display Applied Techniques  
 Document Estimation Technique  
 Make Automatic  
 Make Interactive  
 Redisplay Target Transformation Panes

Figure 7.2: Estimation Technique Documentation

After the creation of the appropriate objects is complete the system changes to the Automatic Design Section's Meta-Groups Configuration.

**Make Interactive:** This command prepares the transformed constraints for use in an interactive design. The selected constraints of the Transformed Constraints Pane are collected for preparation. Each constraint is checked for dependency upon only two fundamental properties. The valid constraints are then coerced into interactive constraints and added to the Interactive Design Section: Preparation Configuration's Interactive Constraints Pane. The system then changes configuration to the Preparation Configuration.

**Redisplay Target Transformation Panes:** Redisplays the Target Transformation Section Title Pane, Transformed Constraints Pane, Estimation Techniques Selection Pane, and the Target Transformation Section Command Menu.

**Remove all Transformed Constraints:** Deletes all transformed constraints from the Transformed Constraints Pane. Deleting the previous constraints before beginning a new design often avoids confusion.

**Revert Constraints:** Removes the technique application knowledge from each of the selected transformed constraint. The constraint is thus reverted to its original form. A single constraint is often transformed differently depending upon whether it is used in an automatic design or an interactive design. Reverting the constraint after one transformation eliminates the need to reenter it from the Problem Formulation Section.

**Select all Transformed Constraints:** Selects each of the unselected constraint displayed in the Transformed Constraints Pane.

## 7.4 Section Operation

The Target Transformation Section is difficult to understand. This is partially because the representation used for the problem of estimation procedure creation is not the best. It is also partially due to the fact that creating estimation procedures is not a task typically done by the chemical engineer. Section 7.5 discusses some of the recommendations I have for improving the section.

The section provides facilities for accomplishing one task: creating estimation procedures. The concept of an estimation procedure is discussed in Volume 1. In brief an estimation procedure for a physical property,  $PP$ , is a collection of equation oriented and group contribution estimation techniques which allow  $PP$  to be estimated for a molecule given only that molecule's molecular structure.

I explain this task further by showing some of the steps involved in creating an estimation procedure for  $P_{vp}$ . The initial state of the Target Transformation Section is shown in Figure 7.3. Each of the **transformed-constraint** objects contains the **property-constraint** object entered in the Problem Formulation Section. The **transformed-constraint** object displays four attributes:

1. **Label**
2. **Value**
3. **Non-Fund**
4. **Fund**

## Target Transformation Section

Transformed Constraints	Estimation Techniques Selection
<p><b>Transformed Constraint</b> Label: Cpl Small Value: (&lt; (CPL 294.3) 32.2) Non-Fund: ( C<sub>pl</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Hv Large Value: (&gt; (HV 272.05) 18.4) Non-Fund: ( ΔH<sub>v</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Pvp High Value: (&lt; (PVP 316.4) 14) Non-Fund: ( P<sub>vp</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Pvp Low Value: (&gt; (PVP 272.05) 1.4) Non-Fund: ( P<sub>vp</sub> )</p>	<p><b>Acentric Factor Techniques</b> <math>P_c \rightarrow (F_1 F_2 F_3)</math> - Joback <math>P_c</math> Factor - Equation Oriented Technique <math>P_c^* \rightarrow ()</math> - Joback <math>P_c^*</math> - Group Contribution Technique <math>P_c \rightarrow (P_c^*)</math> - Joback <math>P_c</math> Modification - Equation Oriented Technique <math>V_c \rightarrow ()</math> - Joback <math>V_c</math> - Group Contribution Technique</p> <p><b>Enthalpy Techniques</b> <math>\Delta H_{298} \rightarrow ()</math> - Joback <math>Hf298</math> - Group Contribution Technique <math>\Delta H_{vb} \rightarrow (F_1 F_2 F_3)</math> - Joback <math>H_{vb}</math> Factor - Equation Oriented Technique <math>\Delta H_{vb} \rightarrow ()</math> - Joback <math>H_{vb}</math> - Group Contribution Technique <math>\Delta H_v \rightarrow (\Delta H_{vb} T_c T_{br})</math> - Fish Lielmezs - Equation Oriented Technique <math>\Delta H_v \rightarrow (\Delta H_{vb} T_c T_b)</math> - Watson Relation <math>T_c</math> Biased - Equation Oriented <math>\Delta H_v \rightarrow (\Delta H_{vb} T_{br} T_b)</math> - Watson Relation <math>T_{br}</math> Biased - Equation Oriented</p> <p><b>Factor Techniques</b> <math>F_2 \rightarrow ()</math> - F2 Assumption - Equation Oriented Technique <math>F_3 \rightarrow ()</math> - Joback F3 - Group Contribution Technique</p>
<p>MD Command: <input type="text"/></p>	<p><input type="checkbox"/> Apply Selected Techniques <input type="checkbox"/> Check Dimensionality <input type="checkbox"/> Check Fundamentality <input type="checkbox"/> Deselect All Techniques <input type="checkbox"/> Deselect All Transformed Constraints</p>

Figure 7.3: Initial State of Transformed Constraints

The **Label** and **Value** attributes display respectively the property-constraint's **label** and **value** instance variables values. The **Non-Fund** attribute displays those required physical properties which are not estimated by any group contribution estimation techniques stored in the **transformed-constraint** object. Required physical properties are those occurring in either the constraint value or resulting from the use of equation oriented estimation techniques. Initially the **transformed-constraint** object contains no estimation techniques so all the **Non-Fund** attributes of all the constraints contain the physical properties occurring in their constraint values.

The designer selects estimation techniques which estimate one of the properties occurring in the **Non-Fund** attribute of the transformed constraint. Figure 7.4 shows the transformed constraint *Pvp High* is selected. The **Display Applicable Techniques** command displays all the estimation techniques which estimate any of the physical properties in the **Non-Fund** attribute. In the case of the *Pvp High* constraint estimation techniques which estimation  $P_{vp}$  are displayed.

The designer selects and applies the Riedel-Plank-Miller  $T_c$  Biased Equation Oriented Estimation Technique. Figure 7.5 shows the system after the technique was applied. Figure 7.5 shows that the non-fundamental properties have changed from  $P_{vp}$  to  $T_b$ ,  $T_c$ , and  $P_c$ . This transformation corresponds to the mapping:

$$P_{vp} \rightarrow (T_b, T_c, P_c)$$

denoted by the estimation technique.

Reapplying the **Display Applicable Techniques** command now gives a new set of applicable estimation techniques. Figure 7.6 shows the applicable techniques. Se-

## Target Transformation Section

Transformed Constraints      Estimation Techniques Selection

**Transformed Constraint**  
Label: Cpl Small  
Value: (< (CPL 294.3) 32.2)  
Non-Fund: ( C<sub>pl</sub> )  
Fund: ( None )

**Transformed Constraint**  
Label: Hv Large  
Value: (> (HV 272.05) 18.4)  
Non-Fund: ( ΔH<sub>v</sub> )  
Fund: ( None )

**Transformed Constraint**  
Label: Pvp High  
Value: (< (PVP 316.0) 18)  
Non-Fund: ( P<sub>pvp</sub> )  
Fund: ( None )

**Transformed Constraint**  
Label: Pvp Low  
Value: (> (PVP 272.05) 1.4)  
Non-Fund: ( P<sub>pvp</sub> )

**Vapor Pressure Techniques**  
 $P_{vp} \rightarrow (T_b \ T_{br} \ P_c) - \text{Riedel Plank Miller Tbr Biased - Equation Orient}$   
 $P_{vp} \rightarrow (T_b \ T_{br} \ P_c) - \text{Riedel Plank Miller Tbr Biased Estimation Oriented}$

MD Command:

Deselect All Techniques  
 Deselect All Transformed Constraints  
 Display All Techniques  
 Display Applicable Techniques  
 Display Applied Techniques

Figure 7.4: Applicable Estimation Techniques for the *Pvp High* Transformed Constraint

## Target Transformation Section

Transformed Constraints      Estimation Techniques Selection

**Transformed Constraint**  
Label: Cpl Small  
Value: (< (CPL 294.3) 32.2)  
Non-Fund: ( Cpl )  
Fund: ( None )

**Transformed Constraint**  
Label: Hv Large  
Value: (> (HV 272.05) 18.4)  
Non-Fund: ( ΔHv )  
Fund: ( None )

**Transformed Constraint**  
Label: Cpl High  
Value: (< (CPL 316.4) 14)  
Non-Fund: ( Tp, Tf, Pv )  
Fund: ( None )

**Transformed Constraint**  
Label: Pvp Low  
Value: (> (PVP 272.05) 1.4)  
Non-Fund: ( Pv )

**Vapor Pressure Techniques**  
 $P_{vp} \rightarrow (T_b, T_{br}, P_c) - \text{Riedel Plank Miller Tbr Biased - Equation Orient}$   
 $P_{vp} \rightarrow (T_b, T_c, P_c) - \text{Riedel Plank Miller Tc Biased - Equation Orient}$

MD Command: Apply Selected Techniques  
MD Command:

Apply Selected Techniques  
 Check Dimensionality  
 Check Fundamentality  
 Deselect All Techniques  
 Deselect All Transformed Constraints

Figure 7.5: Applying the Riedel Plank Miller Technique to Pvp High

lecting and applying Joback's  $T_b$  Group Contribution Technique cause  $T_b$  to become a fundamental physical property – one estimated by a group contribution estimation technique. Figure 7.7 shows the updated constraint object.  $T_b$  was removed from the Non-Fund: list and added to the Fund: list.

The selection and application of estimation techniques continue until all transformed constraints can be estimated by group contribution techniques. This corresponds to all constraints having no non-fundamental physical properties. Figure 7.8 shows the result of applying an appropriate set of estimation techniques. All transformed constraints contain no non-fundamental physical properties.

## 7.5 Section Discussion

The current implementation of the Target Transformation Section poorly represents the underlying problem of estimation procedure formation. The basic task being done by the designer is the traversing of a decision tree choosing the estimation techniques he or she desires for each of the required physical properties. Representing the process as the traversing of a decision tree would make the section much easier to use.

The Evaluation Section's Specifications Configuration is what the next generation of the Target Transformation should look like. The required physical properties would denote the top nodes of the decision tree. The estimation techniques known to the system would be denoted as the children of these property nodes. As the designer selects the desired estimation techniques the required physical properties are updated. Group contribution estimation techniques have no required properties. Choosing stops

**Target Transformation Section**

**Transformed Constraints**

**Transformed Constraint**  
**Label:** Cpl Small  
**Value:** ( $<$  (CPL 294.3) 32.2)  
**Non-Fund:** (  $C_{pl}$  )  
**Fund:** ( None )

**Transformed Constraint**  
**Label:** Hv Large  
**Value:** ( $>$  (HV 272.05) 18.4)  
**Non-Fund:** (  $\Delta H_v$  )  
**Fund:** ( None )

**Transformed Constraint**  
**Label:** Htr High  
**Value:** ( $<$  (HTR 318.4) 14)  
**Non-Fund:** (  $T_c, T_b, P_c$  )  
**Fund:** ( None )

**Transformed Constraint**  
**Label:** Pvp Low  
**Value:** ( $>$  (PVP 272.05) 1.4)  
**Non-Fund:** (  $P_{vp}$  )

**Estimation Techniques Selection**

**Critical Properties Techniques**  
 $P_c \rightarrow (F_1 F_2 F_3)$  - Joback  $P_c$  Factor - Equation Oriented Technique  
 $P_c \rightarrow (P_c^*)$  - Joback  $P_c$  Modification - Equation Oriented Technique

**Temperature Techniques**  
 $T_b \rightarrow (F_1 F_2 F_3)$  - Joback  $T_b$  Factor - Equation Oriented Technique  
 $T_b \rightarrow ()$  - Joback  $T_b$  - Group Contribution Technique  
 $T_c \rightarrow (F_1 F_2 F_3)$  - Joback  $T_c$  Factor - Equation Oriented Technique  
 $T_c \rightarrow (T_b T_{br})$  -  $T_c$  Definition - Equation Oriented Technique

MD Command: Display Applicable Techniques  
 MD Command:  
 MD Command: **Display Applicable Techniques**

Deselect All Transformed Constraints  
 Display All Techniques  
 **Display Applicable Techniques**  
 Display Applied Techniques  
 Document Estimation Technique

Figure 7.6: Displaying Applicable Estimation Techniques

## Target Transformation Section

Transformed Constraints	Estimation Techniques Selection
<p><b>Transformed Constraint</b> Label: Cpl Small Value: (&lt; (CPL 294.3) 32.2) Non-Fund: ( C<sub>pl</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Hv Large Value: (&gt; (HV 272.05) 18.4) Non-Fund: ( ΔH<sub>v</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Tpv High Value: (&lt; (TPV 318.4) 14) Non-Fund: ( C<sub>tpv</sub> ) Fund: ( None )</p> <p><b>Transformed Constraint</b> Label: Ppv Low Value: (&gt; (PVP 272.05) 1.4) Non-Fund: ( P<sub>pv</sub> )</p>	<p><b>Critical Properties Techniques</b> <math>P_c \rightarrow (F_1 F_2 F_3)</math> - Joback P<sub>c</sub> Factor - Equation Oriented Technique <math>P_c \rightarrow (P_c^*)</math> - Joback P<sub>c</sub> Modification - Equation Oriented Technique</p> <p><b>Temperature Techniques</b> <math>T_b \rightarrow (F_1 F_2 F_3)</math> - Joback T<sub>b</sub> Factor - Equation Oriented Technique <math>T_b \rightarrow ()</math> - Joback T<sub>br</sub> - Group Contribution Technique <math>T_c \rightarrow (F_1 F_2 F_3)</math> - Joback T<sub>c</sub> Factor - Equation Oriented Technique <math>T_c \rightarrow (T_b T_{br})</math> - T<sub>c</sub> Definition - Equation Oriented Technique</p>
<p>MD Command: Apply Selected Techniques MD Command: <input type="text"/></p>	<p><b>Apply Selected Techniques</b> Check Dimensionality Check Fundamentality Deselect All Techniques Deselect All Transformed Constraints</p>

Figure 7.7: Applying a Group Contribution Technique

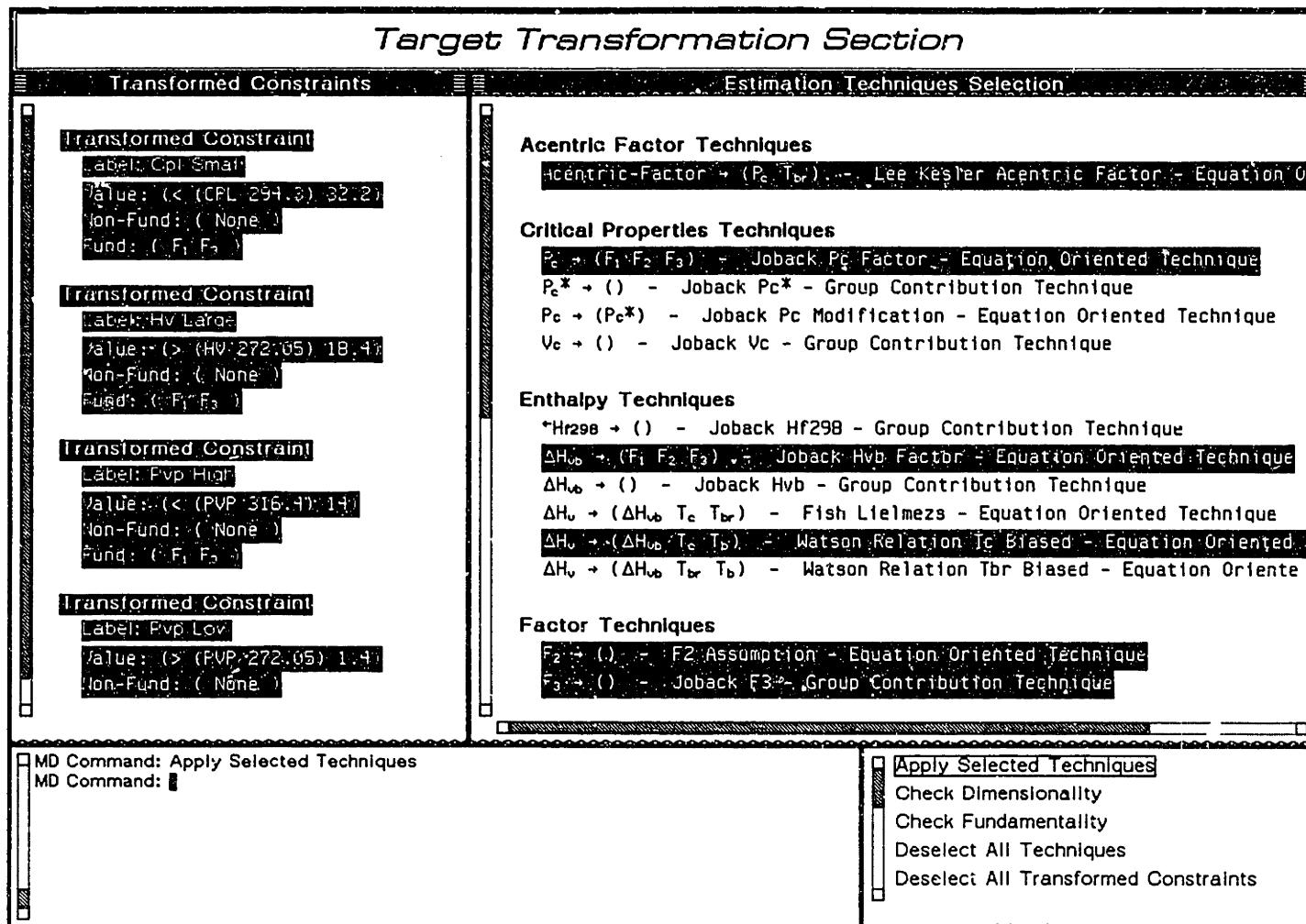


Figure 7.8: Complete Transformation to Factor Dependency

when estimation techniques have been specified for all the required physical properties.

A further enhancement to the section would be the incorporation of "default" estimation techniques. By specifying that a particular physical property should be estimated by technique A when used in an interactive design and technique B when used in an automatic design, the Target Transformation Section could be altogether eliminated. Estimation procedures could still be created by designers who have differences with the system's defaults or who enter new physical properties or estimation techniques.

I find the Target Transformation Section necessary but awkward as it is currently implemented. The two recommendations given above could relegate the section to the background of the system not used by the typical designer. Some thought should be given to the possibility of eliminating it entirely.

## 7.6 Example Usage

The goal of the Target Transformation Section is to provide facilities for the construction of estimation procedures for each of the physical properties occurring in the design constraints. Estimation procedures are discussed in Volume 1. I demonstrate two transformations. The first transforms constraints for use in interactive design. Emphasis is placed on forming estimation procedures which require two fundamental physical properties. The second transforms constraints for use in automatic design. Emphasis is placed on forming estimation procedures which are accurate regardless of the number of fundamental physical properties required.

We begin with the constraints entered in the Problem Formulation Section example

usage.

**Action 7.1** *Mouse right on a empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 7.2** *Mouse left on the Problem Formulation Section Constraints Configuration.*

The system changes configuration to the Problem Formulation Section.

Only the selected constraints are transformed.

**Action 7.3** *Mouse left on the Select All Constraints command.*

All the constraints displayed in the Constraints Pane are selected.

**Action 7.4** *Mouse left on the Transform Constraints command.*

The system coerces each of the constraints into a transformed constraint. These transformed constraints are added to the Target Transformation Section's Transformed Constraints Pane. The system changes to the Target Transformation Section.

## Transformation for Interactive Design

**Action 7.5** *Mouse left on the Select All Transformed Constraints command.*

All the transformed constraints displayed in the Transformed Constraints Pane are selected.

**Action 7.6** *Mouse h-sh-left on the Acentric-Factor ( $P_c T_{br}$ ) - Lee Kesler Acentric Factor – Equation Oriented Technique.*

**Action 7.7** *Mouse h-sh-left on the  $P_c \rightarrow (F_1 F_2 F_3)$  – Joback  $P_c$  Factor – Equation Oriented Technique.*

**Action 7.8** *Mouse h-sh-left on the  $H_{vb} \rightarrow (F_1 F_2 F_3)$  – Joback  $H_{vb}$  Factor – Equation Oriented Technique.*

**Action 7.9** *Mouse h-sh-left on the  $H_v \rightarrow (\Delta H_{vb} T_c T_b)$  – Watson Relation  $T_c$  Biased – Equation Oriented Technique.*

**Action 7.10** *Mouse h-sh-left on the  $F2 \rightarrow ()$  –  $F2$  Assumption – Equation Oriented Technique.*

**Action 7.11** *Mouse h-sh-left on the  $F3 \rightarrow ()$  – Joback  $F3$  – Group Contribution Technique.*

**Action 7.12** *Mouse h-sh-left on the  $F1 \rightarrow ()$  – Joback  $F1$  – Group Contribution Technique.*

**Action 7.13** *Mouse h-sh-left on the  $C_{pl} \rightarrow (C_{pv} \text{ omega } T_c)$  – Rowlinson  $C_{pl}$  – Equation Oriented Technique.*

**Action 7.14** *Mouse h-sh-left on the  $C_{pv} \rightarrow (F_1 F_2 F_3)$  – Joback  $C_{pv}$  298 Factor – Equation Oriented Technique.*

**Action 7.15** *Mouse h-sh-left on the  $T_c \rightarrow (F_1 F_2 F_3)$  – Joback  $T_c$  Factor – Equation Oriented Technique.*

**Action 7.16** *Mouse h-sh-left on the  $T_{br} \rightarrow (T_b T_c)$  –  $T_{br}$  Definition – Equation Oriented Technique.*

**Action 7.17** *Mouse h-sh-left on the Tb  $\rightarrow (F_1 F_2 F_3)$  – Joback Tb Factor – Equation Oriented Technique.*

**Action 7.18** *Mouse h-sh-left on the Pvp  $\rightarrow (Tb Tc Pc)$  – Riedel Plank Miller Tc Biased – Equation Oriented Technique.*

**Action 7.19** *Mouse left on the Apply Selected Techniques command.*

All transformed constraints contain only two fundamental physical properties and no non-fundamental physical properties. Two commands assist in confirming these observations.

**Action 7.20** *Mouse left on the Check Fundamentality command.*

The system checks each of the selected constraints ensuring that each does not contain any non-fundamental physical properties. The system reports that all constraints were verified:

**All Selected Constraints have been Reduced to Fundamental Properties.**

**Action 7.21** *Mouse left on the Check Dimensionality command.*

The system checks each of the selected constraints ensuring that the number of fundamental physical properties is less than or equal to three. The system reports that all constraints were verified.

**All Selected Constraints are of Dimension 3 or Less.**

## Transformation for Automatic Design

The estimation procedures used in the automatic design procedure are developed with an emphasis on accuracy. The dimensionality of the final constraints is not important.

We first deselect all the previously selected estimation techniques.

### Action 7.22 *Mouse left on the Deselect All Techniques command.*

The transformed constraints have all been reduced to fundamental properties using the estimation procedures developed for interactive design. These constraints are all reverted back to their original values.

### Action 7.23 *Mouse left on the Select All Transformed Constraints command.*

### Action 7.24 *Mouse left on the Revert Constraints command.*

We are ready to begin developing new estimation procedures for automatic design.

### Action 7.25 *Mouse h-sh-left on the Acentric-Factor $\rightarrow (Pc\ Tbr)$ – Lee Kesler Acentric Factor – Equation Oriented Technique.*

### Action 7.26 *Mouse h-sh-left on the $Pc^*$ $\rightarrow ()$ – Joback $Pc^*$ – Group Contribution Technique.*

### Action 7.27 *Mouse h-sh-left on the $Pc$ $\rightarrow (Pc^*)$ – Joback $Pc$ Modification – Equation Oriented Technique.*

### Action 7.28 *Mouse h-sh-left on the $Hvb$ $\rightarrow ()$ – Joback $Hvb$ – Group Contribution Technique.*

**Action 7.29** *Mouse h-sh-left on the Hv  $\rightarrow$  (Hvb Tbr Tb) – Watson Relation Tbr Biased – Equation Oriented Technique.*

**Action 7.30** *Mouse h-sh-left on the C<sub>pv</sub><sup>o</sup>  $\rightarrow$  (CpvA, CpvB, CpvC, CpvD) – Cpv Cubic – Equation Oriented Technique.*

**Action 7.31** *Mouse h-sh-left on the CpvD  $\rightarrow$  () – Joback CpvD – Group Contribution Technique.*

**Action 7.32** *Mouse h-sh-left on the CpvC  $\rightarrow$  () – Joback CpvC – Group Contribution Technique.*

**Action 7.33** *Mouse h-sh-left on the CpvB  $\rightarrow$  () – Joback CpvB – Group Contribution Technique.*

**Action 7.34** *Mouse h-sh-left on the CpvA  $\rightarrow$  () – Joback CpvA – Group Contribution Technique.*

**Action 7.35** *Mouse h-sh-left on the Cpl  $\rightarrow$  (Cpv omega Tc) – Rowlinson Cpl – Equation Oriented Technique.*

**Action 7.36** *Mouse h-sh-left on the Tc  $\rightarrow$  (Tb Tbr) – Tc Definition – Equation Oriented Technique.*

**Action 7.37** *Mouse h-sh-left on the Tbr\*  $\rightarrow$  () – Joback Tbr\* – Group Contribution Technique.*

**Action 7.38** *Mouse h-sh-left on the Tbr  $\rightarrow$  (Tbr\*) – Joback Tbr Modification – Equation Oriented Technique.*

**Action 7.39** *Mouse h-sh-left on the Tb → () – Joback Tb – Group Contribution Technique.*

**Action 7.40** *Mouse h-sh-left on the Pvp → (Tb Tbr Pc) – Riedel Plank Miller Tbr Biased – Equation Oriented Technique.*

**Action 7.41** *Mouse left on the Apply Selected Techniques command.*

Check that all transformed constraints are reduced to fundamental properties.

**Action 7.42** *Mouse left on the Check Fundamentality command.*

The system confirms that all constraints are verified.

**All Selected Constraints have been Reduced to Fundamental Properties.**

# Chapter 8

## Automatic Design Section

The Automatic Design Section implements the automatic design methodology described in Volume 1. The primary tasks of the section are to expand and divide meta-groups, form meta-molecules from these meta-groups, and prune these meta-molecules using structural and physical property constraints. These tasks are divided over two configurations: the Meta-Groups Configuration and the Meta-Molecules Configuration.

### 8.1 Section Layout

The screen layout of the Meta-Groups Configuration is shown in Figure 8.1. The screen real estate is used by five panes:

**Automatic Design Section Meta-Groups Configuration Title Pane:** Displays the title of the Meta-Groups Configuration.

## Automatic Design Section: Meta-Groups

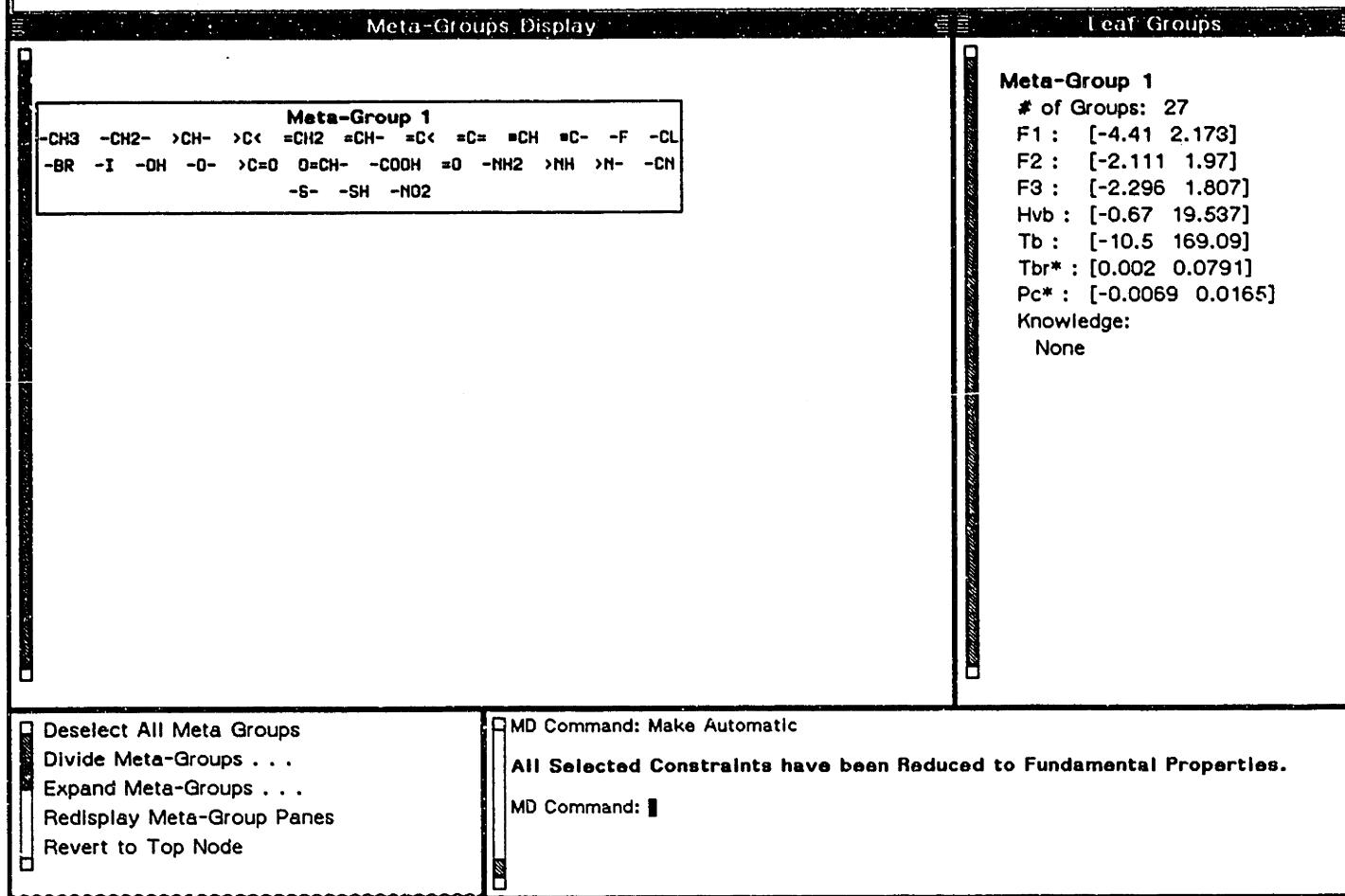


Figure 8.1: Meta-Groups Configuration Screen

**Meta-Groups Display Pane:** Displays all the meta-groups created by the system in this design. The meta-groups are displayed in a tree structure showing the ancestry of all the groups. Each node of the tree is a meta-group displayed as a box containing the name of the meta-group and the set of groups it contains.

The origin of the tree is the top meta-group which contains all the consistent groups used in the design. Meta-group 1 in Figure 8.1 is a top meta-group. When a meta-group is divided or expanded the tree is updated. The Meta-Groups Display Pane in Figure 8.2 displays the updated tree after meta-group 1 was expanded by global valence.

**Leaf Groups Pane:** Displays the leaf nodes of the meta-group tree displayed in the Meta-Groups Display Pane. These are the meta-groups currently used in any meta-molecule. Each leaf group is displayed listing its name, number of groups, knowledge, and interval values for each of the fundamental properties being used in the design.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Automatic Design Section Meta-Groups Configuration Commands Menu:** The command menu containing commands relevant to the Meta-Groups Configuration.

The screen layout of the Meta-Molecules Configuration is shown in Figure 8.3. The screen real estate is used by six panes:

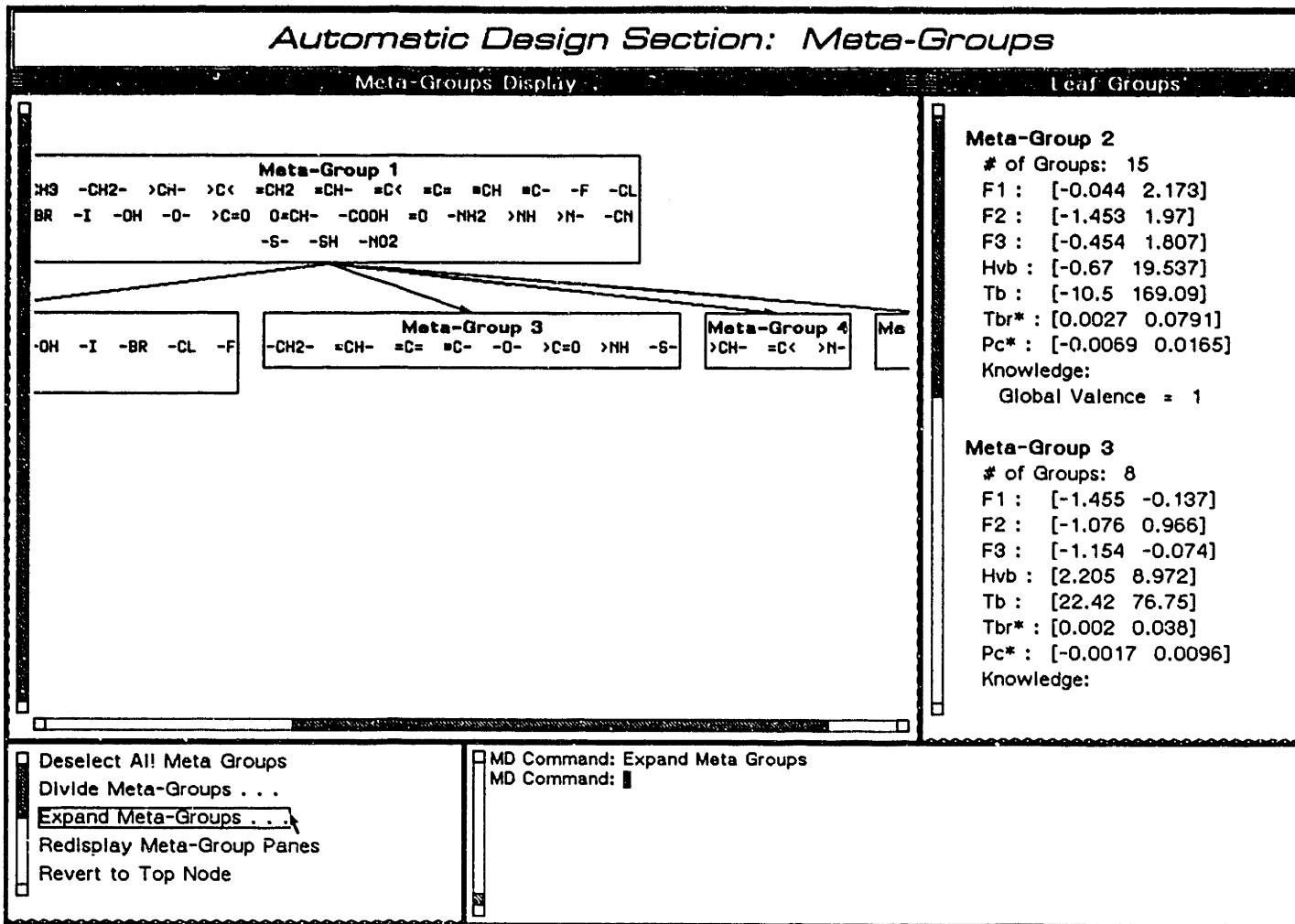


Figure 8.2: Meta-Group 1 Expanded by Global Valence

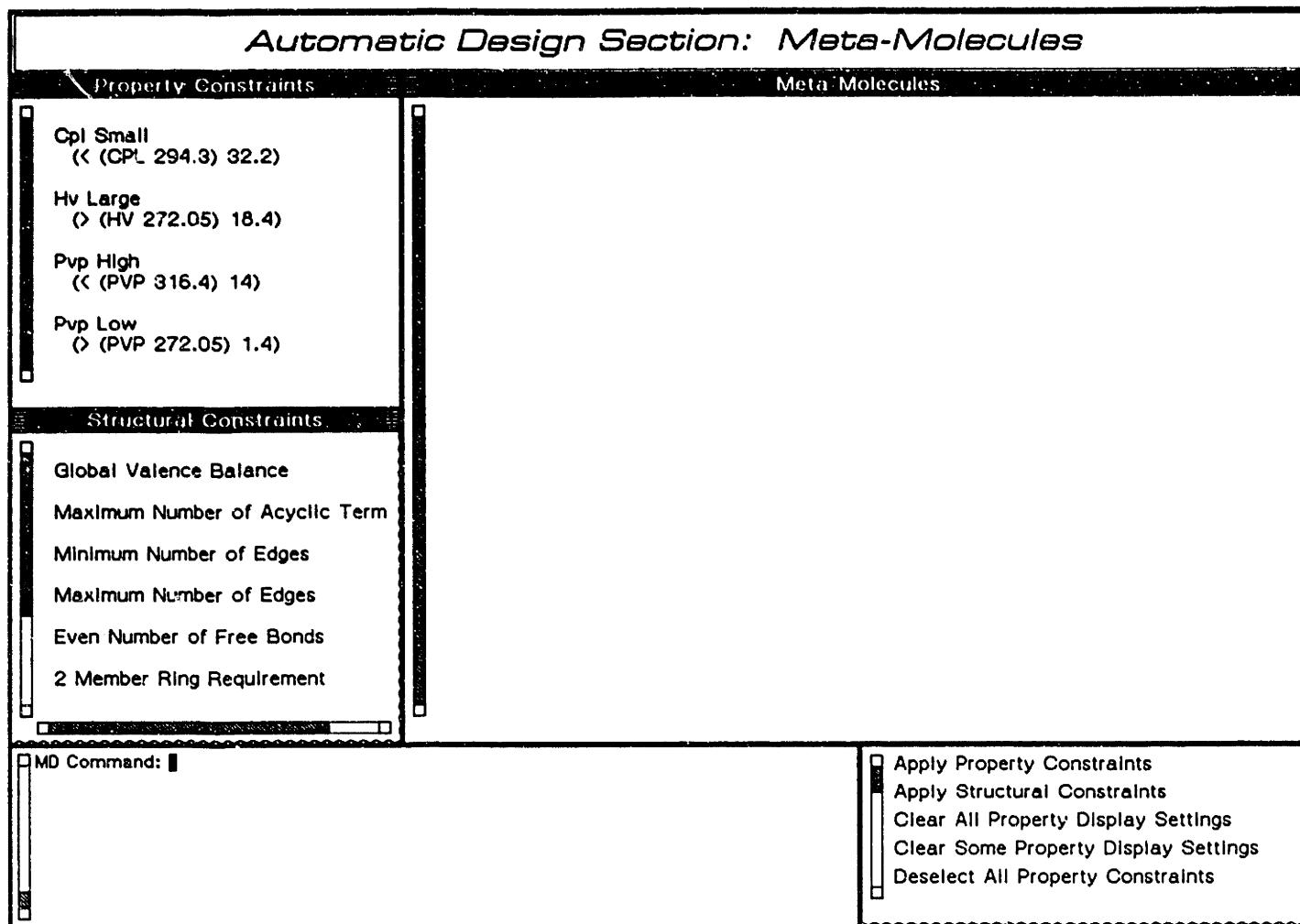


Figure 8.3: Meta-Molecules Configuration Screen

**Automatic Design Section Meta-Molecules Configuration Title Pane:** Displays the title of the Meta-Molecules Configuration.

**Property Constraints Pane:** Displays the Physical Property Constraints available for pruning meta-molecules. The Target Transformation Section's **Make Automatic** command coerced the selected transformed constraints into automatic constraints and placed them in this pane.

**Structural Constraints Pane:** Displays the structural constraints available for pruning meta-molecules. Eight structural constraints are currently known to the system:

1. Global Valence Balance
2. Minimum Number of Edges
3. Maximum Number of Edges
4. Even Number of Ring Bonds
5. 2 Member Ring Requirement
6. 3 Member Ring Requirement
7. Necessity of Non-Mixed Groups
8. Existence

These constraints are explained in Volume 1.

**Meta Molecules Pane:** Displays the meta-molecules involved in the current design. Meta-molecules are presented as lists of their meta-group occurrences. Displaying many meta-molecules takes considerable time. Whenever the number of meta-molecules exceeds 250 the system prompts the designer for confirmation that he or she wants all

the meta-molecules displayed. If the designer answers “No” a message is displayed in the Meta Molecules Pane noting the number of meta-molecules present.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Automatic Design Section Meta-Molecules Configuration Commands Menu:** The command menu containing commands relevant to the Meta-Molecules Configuration.

## 8.2 Section Operation

The Automatic Design Section produces meta-molecules which are pruned by property and structural constraints. The major component of a meta-molecule is a list of meta-group occurrences. The occurrence list and the structural characteristics of each meta-group is used by structural constraints to test for feasibility. The occurrence list and the meta-contributions of each meta-group is used to calculate meta-properties for each meta-molecule. These meta-properties are used to test for satisfaction of property constraints.

The Automatic Design Section is typically reached from the Target Transformation Section. The Target Transformation Section’s **Make Automatic** command finds a consistent set of groups from the chosen group contribution techniques. These groups are used to form a meta-group which is added to the Meta-Groups Configuration’s

Meta-Groups Display Pane. The selected transformed constraints are coerced into automatic property constraints and added to the Meta-Molecules Configuration's Property Constraints Pane. Finally the system changes to the Meta-Groups Configuration.

The designer begins an automatic design by entering an initial set of meta-molecules. The **Input Meta-Molecules** command prompts for an upper and lower bound on the number of occurrences used to form new meta-molecules. This command then creates all appropriate meta-molecules and adds them to the Meta-Molecules Pane. Figure 8.4 shows nine meta-molecules created by the **Input Meta-Molecules** command.

To use structural constraints the system must know the structural characteristics of each meta-group. Knowledge of meta-group structure is identified whenever meta-groups are expanded. The **Expand Meta-Groups** command displays a set of characteristics which can be used to expand our top meta-group. Figure 8.5 shows the result of expanding our top meta-group by the global valence structural characteristic.

## 8.3 Meta-Groups Configuration Objects

The two major objects used in the Data Configuration are:

1. **Meta-Group-Object**
2. **Leaf-Node-Object**

The definitions for these objects and their associated functions are in the file:

```
molecular-design:automatic-design-section;objects.lisp.
```

I discuss the instance variables and important functionality for both of these objects.

**Meta-groups** have three major instance variables:

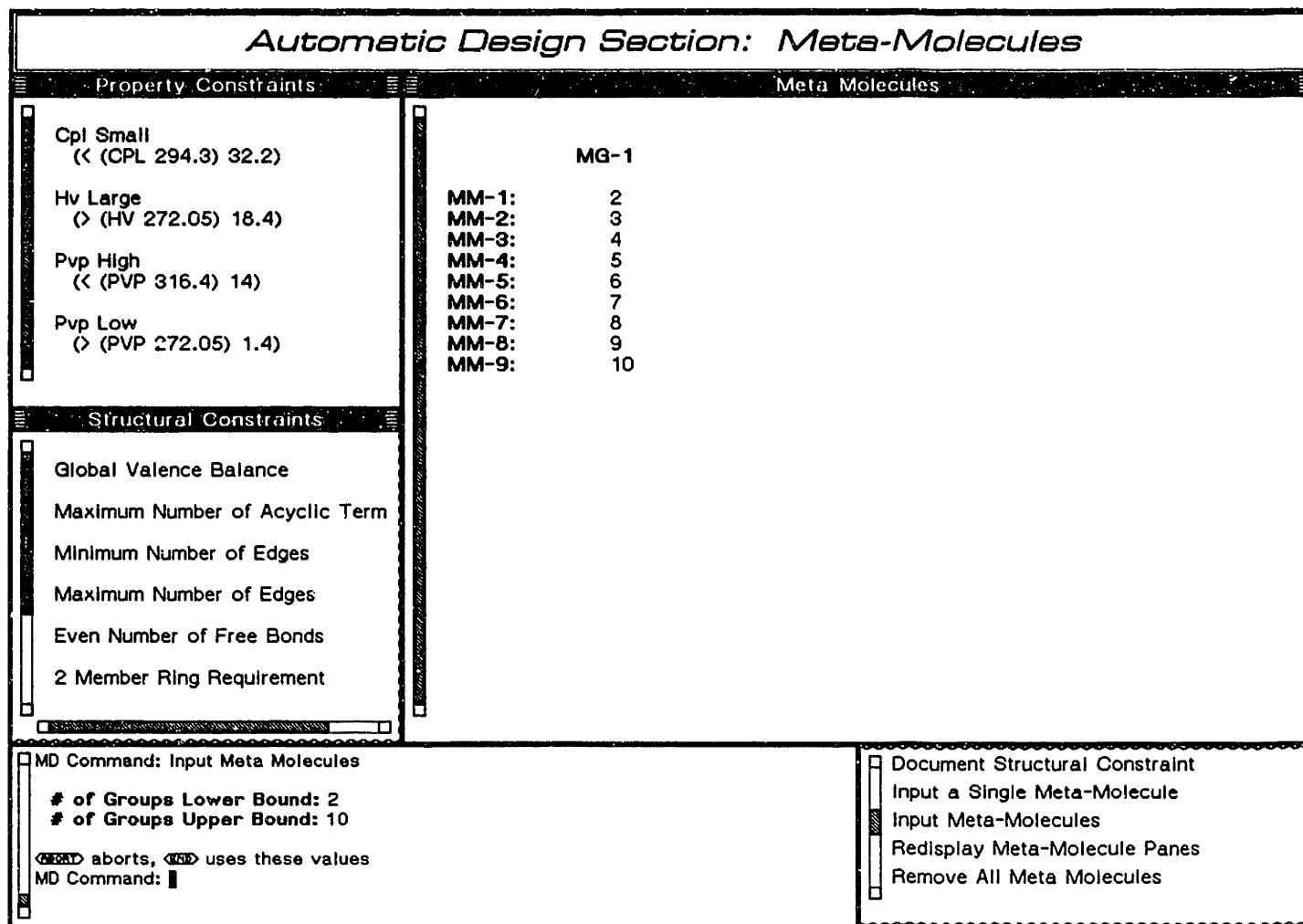


Figure 8.4: Initial Set of Meta-Molecules

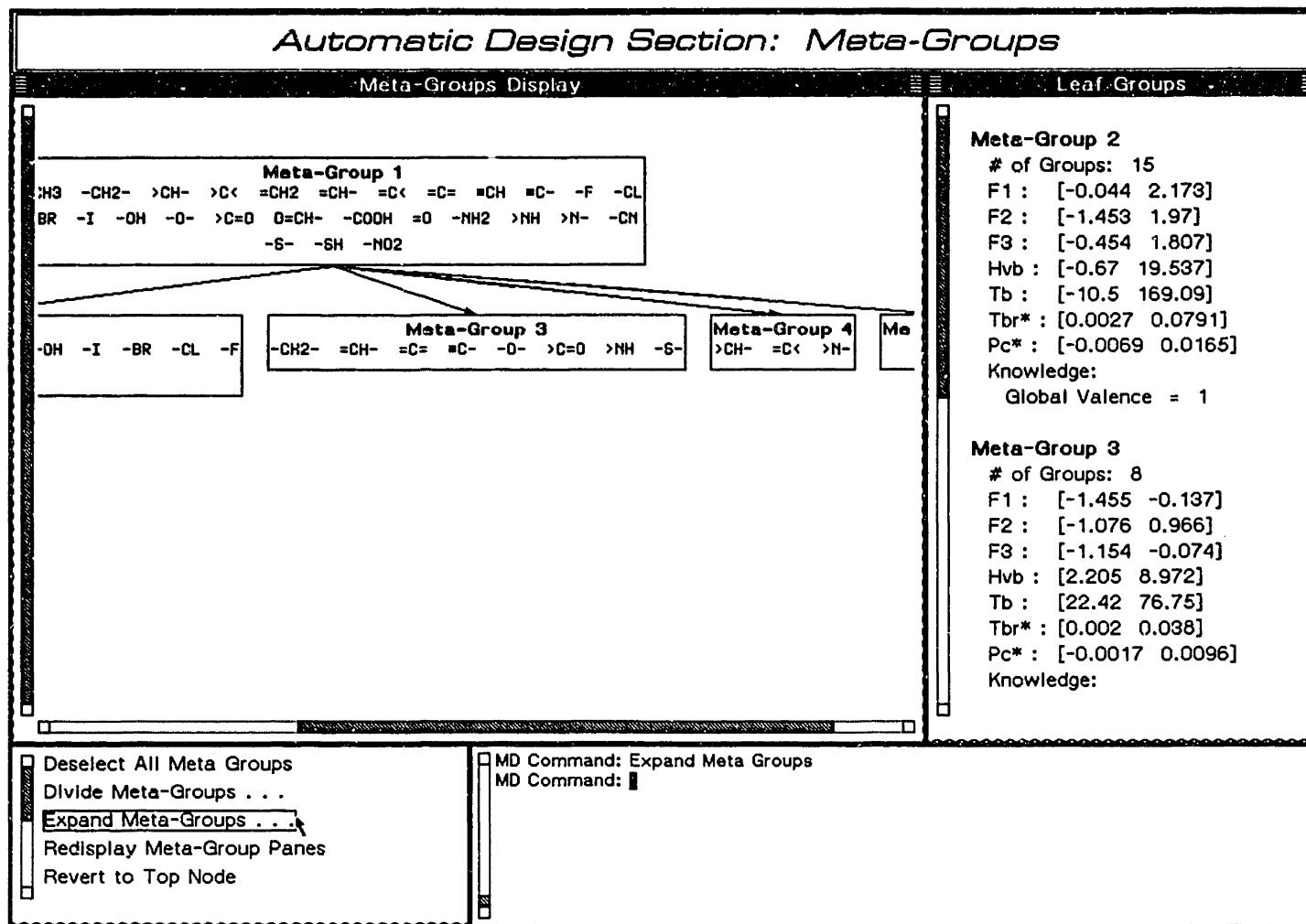


Figure 8.5: Expansion by the Global Valence Structural Characteristic

1. **groups**
2. **knowledge-alist**
3. **meta-contributions**
4. **children**

The major tasks of a **meta-group** are to keep track of groups and meta-contributions.

The **groups** instance variable contains the list of all the groups contained in the meta-group. Contributions and molecular knowledge are extracted from these groups.

The **knowledge-alist** instance variable keeps track of the information known about the groups. Knowledge is the global valence, bond type, ring class, etc. This knowledge is used by the structural constraints to prune structurally infeasible meta-molecules.

Whenever a **meta-group** is formed a **meta-contribution** is formed for each of the fundamental physical properties required in the design. The **meta-contributions** instance variable keeps a list of these **meta-contributions**. The **children** instance variable keeps track of the parent-child relationships of the meta-groups. These relationships are used for display.

**Leaf-groups** are used to simplify the display of meta-groups. The Meta-Groups Display Pane displays meta-groups in a tree originating from the initial meta-group. This is a very pretty display but is limited in the amount of information it can present. **Leaf-groups** display the information contained in meta-groups in a “tabular” format. The number of groups a meta-group contains, the meta-contribution for each of the fundamental physical properties, and the knowledge contained are all displayed.

## 8.4 Meta-Molecules Configuration Objects

The three major objects of the meta-molecules configuration are:

1. **Meta-Molecule-Object**
2. **Automatic-Constraint-Object**
3. **Structural-Constraint-Object**

The definitions for these objects and their associated functions are in the file:

`molecular-design:automatic-design-section;objects.lisp.`

The instance variables and important functionality for each of these objects is discussed.

**Meta-molecules** represent the major object in the automatic design section. Each meta-molecule contains one major instance variable – the **occurrences-list**. The **occurrences-list** contains the number of occurrences of each of the current meta-groups. **Automatic-constraint** objects use this information to estimate the interval values for each of the fundamental properties and check each meta-molecule for satisfaction of the property constraints. **Structural-constraint** objects use the occurrences of meta-groups in a meta-molecule to check feasibility constraints.

## 8.5 Meta-Groups Configuration Commands

The following commands are listed in the Command Menu of the Automatic Design Section. The definition for the **Expand Meta-Groups ...** command is in the file:

`molecular-design:automatic-design-section;commands-expansion.lisp.`

The definition for the **Select Meta-Groups ...** command is in the file:

`molecular-design:automatic-design-section;commands-selection.lisp.`

The definitions of the remaining commands are in the file:

`molecular-design:automatic-design-section;commands-meta-groups.lisp.`

**Deselect all Meta Groups:** Leaf groups are selected for division. This command deselects all the selected `leaf-groups` displayed in the Leaf Groups Pane.

**Divide Meta-Groups . . .:** Dividing a `meta-group` creates two or more meta-groups containing subsets of the divided meta-group's groups. Division is applied only to selected leaf groups. Several methods of division are available:

**Divide Manually:** The groups contained in the selected meta-groups are displayed. The designer can manually divide these groups into any two subsets. The two subsets are used to form two new meta-groups.

**Divide in Half:** Divides the set of groups in half.

**Divide in Half wrt Property:** Divides the set of groups in half after sorting with respect to the chosen property's contributions.

**Divide in Half wrt Property Contributions:** Divides the set of groups based upon the midpoint of the chosen property's contributions.

**Divide by Largest Gap wrt Property:** The groups are first sorted with respect to the chosen property's contributions. The largest gap in these contributions is then found. The set of groups is divided into two subsets at the location of the largest gap.

**Divide by Largest % Gap wrt all Properties:** Sorts the contributions for each property and identifies the largest percentage gap. The property which has the largest percentage gap is used to sort and divide the groups into two subsets.

**Divide by Sign of Contributions wrt Property:** Divides the set of groups into two subsets each with a consistent sign for the contributions of a chosen property.

**Divide by Sign of Contributions wrt all Properties:** Divides the set

of groups into two or more subsets. Each group subset has a consistent sign for the contributions for each of the fundamental properties.

**Expand Meta-Groups ...:** Expanding a meta-group creates two or more meta-groups each with a consistent value for molecular knowledge. Expansion applies to all meta-groups present. Several methods of expansion are available:

**Global Valence:** Allocates groups into subsets based upon the group's global valence. The global valence of a group is the number of free bonds it contains regardless of the bond's type.

**Ring Class:** Allocates groups into subsets based upon the group's ring class. There are three ring classes: 1) cyclic; 2) acyclic; 3) mixed. The ring class is determined by the type of free bonds a group contains.

**Bond Type:** Allocates groups into subsets based upon a list of the group's bond types. There are five bond types currently known to the system:

1. `:single-bond`
2. `:double-bond`
3. `:triple-bond`
4. `:ring-single-bond`
5. `:ring-double-bond`

**Bond Valence:** Allocates groups into subsets based upon a list containing the number of each type of bond contained in the group. The bond valence of  $=\text{CH}-$  is `((:double-bond . 1)(:single-bond . 1))`.

**Fully:** Divides the current meta-groups into new meta-groups each containing only a single group.

**Redisplay Meta-Groups Panes:** Redisplays each of the five panes associated with the meta-groups configuration.

**Revert to Top Node:** Restarts the automatic design. All meta-groups except for the top meta-group are removed. All meta-molecules are removed.

**Select All Meta Groups:** Selects all the leaf-groups displayed in the Leaf Groups Pane.

**Select Meta-Groups:** Selects meta-groups using several methods:

**Largest Number of Groups:** Selects the leaf-group which contains the largest number of groups.

**Smallest Number of Groups:** Selects the leaf-group which contains the smallest number of groups.

**Largest Span:** Prompts the designer for a fundamental physical property. Selects the leaf group which has the widest meta-contribution toward this property.

**Smallest Span:** Prompts the designer for a fundamental physical property. Selects the leaf group which has the narrowest meta-contribution toward this property.

**Lower Limit on Number of Groups:** Prompts the designer for a number used to limit the number of groups a leaf-group can contain. Selects all leaf-groups which contain more groups than this number.

**Upper Limit on Number of Groups:** Prompts the designer for a number used to limit the number of groups a leaf-group can contain. Selects all leaf-groups which contain fewer groups than this number.

**Different Signs of Property:** Prompts the designer for a fundamental physical property. Selects all leaf-groups whose meta-contributions span zero.

Selection of meta-groups usually precedes meta-group division.

**Show Meta-Contributions:** Displays the meta-contributions for a chosen group on a window resource exposed over the *Meta-Groups Display Pane*.

**Show Meta-Group Statistics:** Displays various statistics about the meta-groups configuration on a window resource exposed over the *Meta-Groups Display Pane*. The statistics shown are:

1. Total Number of Groups.
2. Total Number of Meta-Groups.
3. Total Number of Leaf Meta-Groups.
4. Total Number of Selected Leaf Meta-Groups.
5. Total Number of Meta-Molecules.

Figure 8.6 shows a screen with the meta-group statistics displayed. Typing any character deactivates the window resource.

**Sort Contributions:** Prompts the designer for the meta-group, the property to sort with respect to, and whether the sorting is to be done in ascending or descending order. When examining meta-contributions for a point of division it is useful to sort contributions. This facilitates identifying gaps in the contributions which could be a good place to divide a meta-group in order to tighten meta-properties.

**Zero Meta-Group Count:** The system keeps track of each meta-group created. A name is given to a newly created meta-group which is the number of the meta-group preceded by the string “Meta-Group”. This number is often useful in keeping track of the number of meta-groups created during a design. After one design is finished, zeroing the count prepares for the next design.

## 8.6 Meta-Molecules Configuration Commands

The following commands are listed in the command menu of the meta-molecules configuration. The definitions of the commands are in the files:

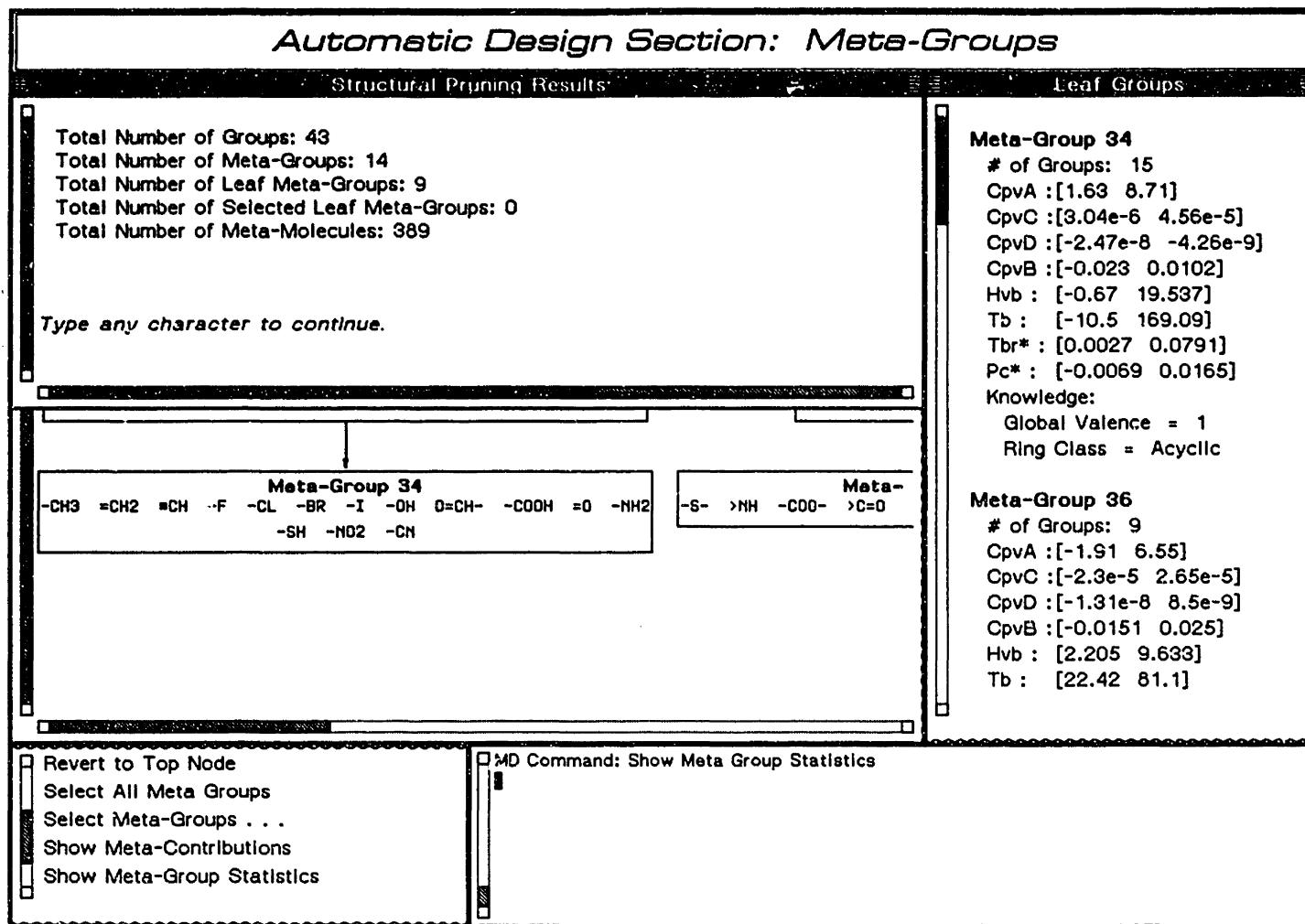


Figure 8.6: Meta-Group Statistics Display

`molecular-design:automatic-design-section;commands-meta-molecules.lisp`

and

`molecular-design:automatic-design-section;constraints.lisp.`

**Apply Property Constraints:** Checks each of the meta-molecules displayed in the Meta Molecules Pane for satisfaction of each of the selected property constraints displayed in the Property Constraints Pane. Meta-molecules which do not satisfy a property constraint are removed from the Meta Molecules Pane.

During the pruning process, the system displays pruning results on a window resource exposed over the Meta Molecules Pane. Figure 8.7 shows the system displaying pruning results.

**Apply Structural Constraints:** Checks each of the meta-molecules displayed in the Meta Molecules Pane for satisfaction of each of the selected structural constraints displayed in the Structural Constraints Pane. Meta-molecules which do not satisfy a structural constraint are removed from the Meta Molecules Pane.

During the pruning process, the system displays pruning results on a window resource exposed of the Meta Molecules Pane. Figure 8.8 shows the results from applying several structural constraints.

**Deselect all Property Constraints:** Deselects all the property constraint objects displayed in the *Property Constraints Pane*.

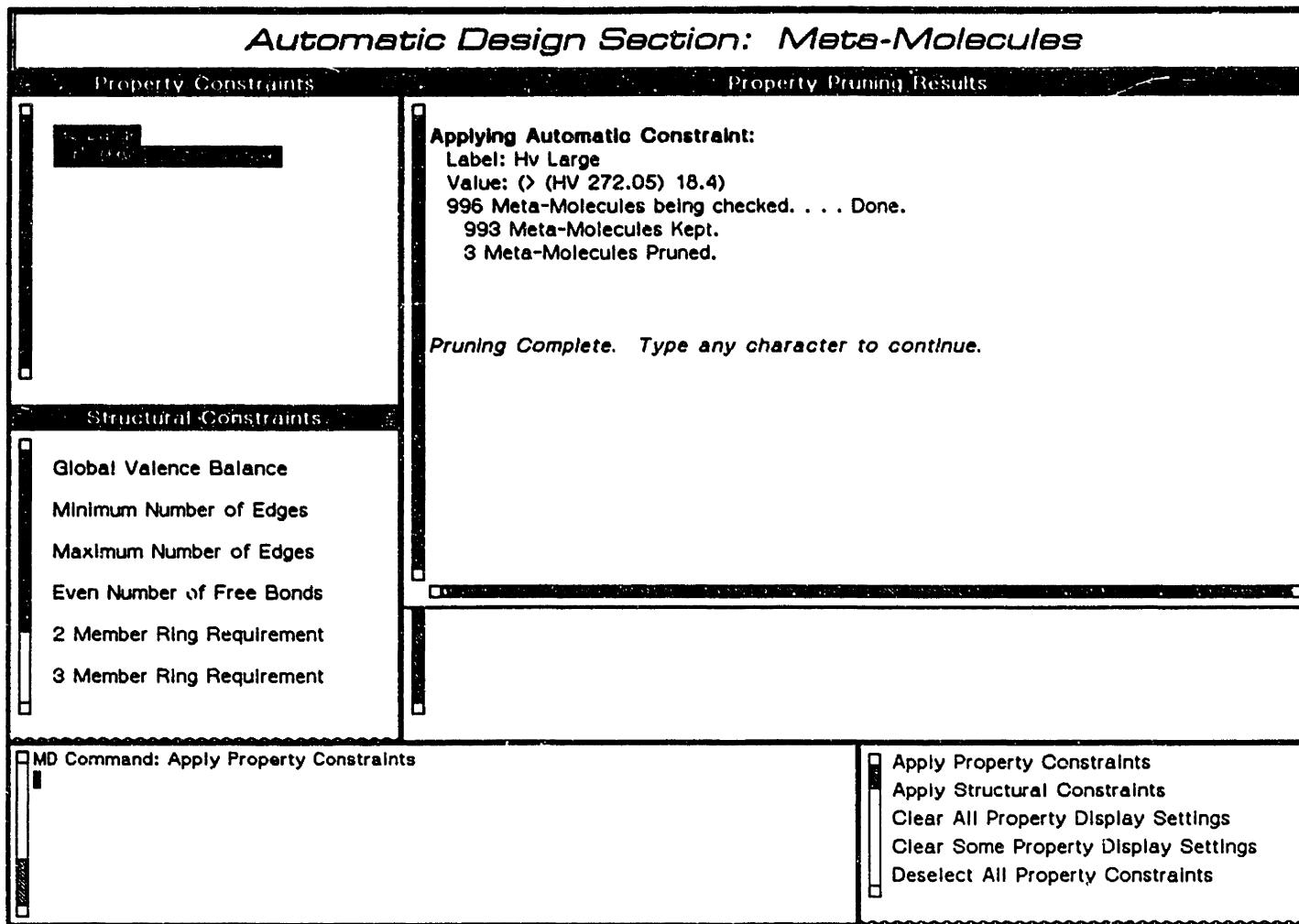


Figure 8.7: Property Constraints Pruning Results

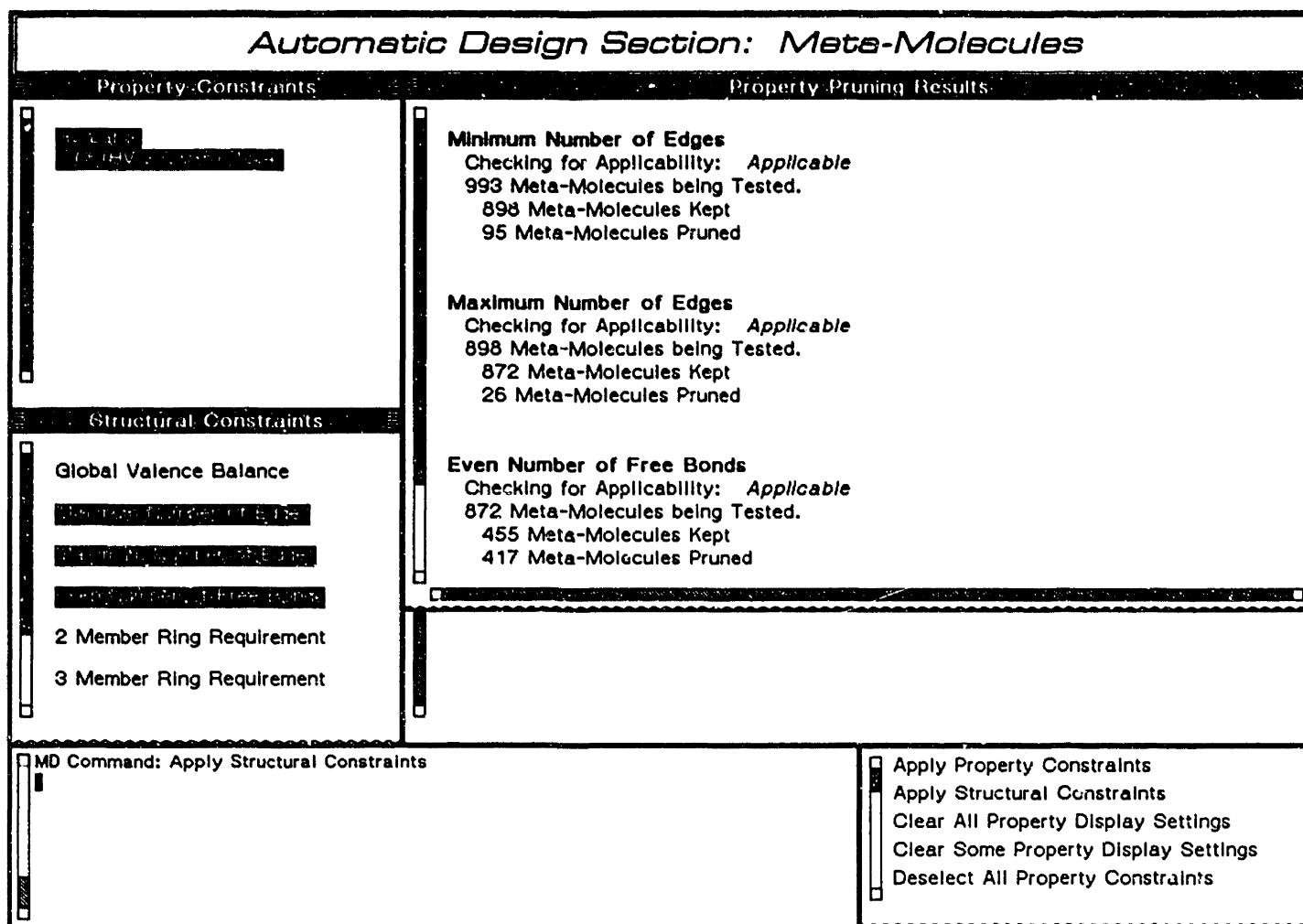


Figure 8.8: Structural Constraints Pruning Results

**Deselect all Structural Constraints:** Deselects all the structural constraint objects displayed in the *Structural Constraints Pane*.

**Document Structural Constraint:** Prompts for a structural constraint. The chosen structural constraint's documentation is shown in a window resource exposed over the Meta Molecules Pane. The documentation usually states what knowledge is required for the constraint to be active.

**Generate Final Report:** Writes the occurrences of meta-groups for each meta-molecule to a file. A typical line in the file would be:

MM-27933 1(-F) 1(-CL) 1(-CH2-)

The command is intended to be used when an automatic design is completed.

**Input a Single Meta-Molecule:** Exposes a menu prompting for the occurrence of each leaf group. Entering a single meta-molecule is useful when checking newly entered structural constraints.

**Input Meta-Molecules:** This command creates and adds meta-molecules to the Meta-Molecules Pane. The major task of this command is to generate the occurrence lists for the new meta-molecules. The system prompts for upper and lower bounds on the number of group occurrences. The system generates occurrence lists equal in length to the number of leaf groups with the total number of occurrences between these limits. For example, if there are currently 4 leaf groups then occurrence lists of length 4 are generated. With input limits of 2 to 3 the occurrence lists generated are:

(2 0 0 0) (0 2 0 0) (0 0 2 0) (0 0 0 2) (1 1 0 0)  
(1 0 1 0) (1 0 0 1) (0 1 1 0) (0 1 0 1) (0 0 1 1)  
(3 0 0 0) (0 3 0 0) (0 0 3 0) (0 0 0 3)  
(2 1 0 0) (2 0 1 0) (2 0 0 1) (0 2 1 0)  
(0 2 0 1) (0 0 2 1) (1 2 0 0) (1 0 2 0)  
(1 0 0 2) (0 1 2 0) (0 1 0 2) (0 0 1 2)  
(1 1 1 0) (1 1 0 1) (1 0 1 1) (0 1 1 1)

All occurrence lists are of length 4. The total number of occurrences in any occurrence list is between 2 and 3.

**Limit Group Occurrences:** Enables the designer to place an upper limit on the number of occurrences of each meta-group any meta-molecule can contain. The system exposes a limit prompting for a limit on each meta-group occurrence. Meta-molecules which have occurrences exceeding these limits are removed from the Meta Molecules Pane.

**Redisplay Meta-Molecules Panes:** Redisplays the five panes of the meta-molecules configuration. The Molecular Design Interaction Pane is not redisplayed.

**Remove all Meta-Molecules:** This command removes all the meta-molecules displayed in the Meta-Molecules Pane. Once removed these meta-molecules are not retrievable.

**Report on Some Meta-Molecules:** Writes information on current leaf-groups and meta-molecules to a file for printing. The information contains the groups contained in each leaf group, the name of each chosen meta-molecule, and the meta-molecule's occurrence list. Generating meta-molecule reports is useful when debugging new structural constraints. Since the system has no molecular display capabilities, the generation of meta-molecule reports is the final result of an automatic design. The information presented in the report is sufficient for the designer to reconstruct the molecules suggested by the system.

The command prompts the designer for the number of meta-molecules to report on. The designer then specifies the criteria for choosing meta-molecules. Meta-molecules are either chosen in the order presented in the *Meta-molecules Pane* or randomly. Finally the system prompts for the file in which to save the information.

**Select all Property Constraints:** Selects all the automatic constraint objects displayed in the *Property Constraints Pane*. Only selected property constraints are used in pruning.

**Select all Structural Constraints:** Selects all the structural constraint objects displayed in the *Structural Constraints Pane*. Only selected structural constraints are used in pruning.

**Select Applicable Constraints:** Selects those structural constraints for which molecular knowledge is available. For example, the Minimum Number of Edges structural constraint would be selected if global valence is known. The command does not know

which constraints have already been applied. It therefore would continue to select the Minimum Number of Edges constraint even after it was applied and additional knowledge was added.

**Show Leaf Groups:** The leaf groups of the Meta-Groups Configuration's Leaf Groups Pane are displayed in a window resource exposed over both the Property Constraints and Structural Constraints Pane. Examining the leaf groups along with their occurrence in meta-molecules provides insight into meta-group expansions.

**Show Meta-Molecule Statistics:** Displays statistics about the Meta-Molecules Configuration on a window resource exposed over the *Meta-Molecules Pane*. The statistics shown are:

1. Total Number of Molecules
2. Total Number of Leaf Groups

**Show Occurrence Statistics:** Displays the following information for each meta-group in a window resource exposed over the Meta Molecules Pane:

1. The number of groups it contains.
2. The average occurrence in the meta-molecules displayed in the Meta Molecules Pane.
3. The maximum occurrence in all of the meta-molecules displayed in the Meta Molecules Pane.
4. The minimum occurrence in all of the meta-molecules displayed in the Meta Molecules Pane.

**Show Some Meta-Molecules:** Displays 10 meta-molecules in a window resource exposed over the Meta Molecules Pane.

**Zero Meta-Molecule Count:** The system keeps a count of all the meta-molecule objects created. This count is used in the naming of meta-molecules. A meta-molecule is named by a string consisting of “MM-” and its number.

This number is often useful in keeping count of the number of meta-molecules created during a design. After one design is finished, zeroing the count prepares for the next design.

## 8.7 Section Discussion

The algorithm I developed demonstrated the feasibility and advantage of using an abstracted generate and test for molecular design. The use of intervals, meta-groups, and structural constraints for the design of molecules still needs exploration. The algorithm used in the automatic design can be improved.

### 8.7.1 Physical Property Pruning

The procedure for pruning meta-molecules with physical property constraints should be significantly changed. Presently there is a great deal of repetitive estimation involved in the pruning. If we applied the two vapor pressure constraints shown in the Property Constraints Pane of Figure 8.4 the system would estimate,  $T_b$ ,  $T_{br}^*$ , and  $P_c^*$  for the *Pvp High* constraint and then reestimate them for the *Pvp Low* constraint. Caching some estimated properties in each meta-molecule would speed pruning.

### 8.7.2 Meta-Molecule Inheritance

The total number of group occurrences in a typical design is 10 or less. The typical number of groups is greater than 20. This implies that a large percentage of the occurrences in a meta-molecule have the value zero. This zero occurrence can be used to eliminate the need to examine many meta-molecules.

Let MM1 and MM2 be meta-molecules having occurrences:

	MG1	MG2	MG3	MG4	MG5	MG6
MM1	2	1	2	0	0	1
MM2	2	1	0	0	0	0

Expanding meta-group MG3 results in the four children meta-molecules:

	MG1	MG2	MG3 <sub>1</sub>	MG3 <sub>2</sub>	MG4	MG5	MG6
MM1 <sub>1</sub>	2	1	2	0	0	0	1
MM1 <sub>2</sub>	2	1	1	1	0	0	1
MM1 <sub>3</sub>	2	1	0	2	0	0	1
MM2 <sub>1</sub>	2	1	0	0	0	0	0

Meta-molecule MM1 creates three children. Meta-molecule MM2 creates a single child. The physical properties of the three children of MM1 are included in the physical property intervals of MM1. However, the physical properties of the child of MM2 are identical to the physical properties of MM2. If a meta-molecule contains no occurrences of a meta-group then expanding that meta-group has no effect on the meta-molecule's physical properties. Since the physical properties are the same then the result of applying property constraints is also the same. If MM2 satisfies a physical property constraint then MM2<sub>1</sub> also satisfies that constraint.

After several expansions a meta-group may occur in only 10% of the meta-molecules being tested. Avoiding retesting meta-molecules which were formed from the expansion of a zero occurrence could thus eliminate the need to test 90% of the meta-molecules present.

Implementing this improvement would require a meta-molecule to keep a record of which property constraints tested it. If the meta-molecule is expanded by a zero occurrence it should pass on to its child this record in addition to the physical properties it possesses. Physical property constraints are only applied to those meta-molecules which have not inherited positive test results from their parents.

## 8.8 Example Usage

An automatic design may take several days to complete. The following example should be able to be completed in less than one day. Much of the time is spent waiting for the system to examine meta-molecules checking to see if they satisfy the input physical property constraints.

### Entering the Property Constraint

We use a single property constraint to demonstrate the automatic design. This constraint is entered in the Problem Formulation Section.

#### Action 8.1 *Mouse right on a empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 8.2** *Mouse left on the Problem Formulation Section Constraints Configuration.*

The system changes configuration to the Problem Formulation Section.

We enter the constraint:

$$P_{vp}(272.05K) > 2.0 \text{ bar.}$$

**Action 8.3** *Mouse left on the Create Constraint command.*

The system displays an accepting-values menu:

**Constraint's Label:** *some value*

**Constraint's Value:** *some value*

**ABORT** aborts, **END** uses these values.

The actual values displayed after the prompts are not important when the menu is initially displayed.

**Action 8.4** *Mouse left on the phrase displayed after the Constraint's Label: prompt.*

*The phrase is replaced by a blinking cursor.*

**Action 8.5** *Type in the label of our constraint: Pvp Low. Press the return key when you complete the entry.*

**Action 8.6** *Mouse left on the phrase displayed after the Constraint's Value: prompt.*

*The phrase is replaced by a blinking cursor.*

**Action 8.7** *Type in the value of our constraint: (> (Pvp 272.05) 2). Press the return key when you complete the entry.*

When both entries are made the constraint is created and added to the Constraints Pane.

**Action 8.8** *Press the end key.*

### Target Transformation

The criteria used when transforming property constraints for automatic design is that of highest accuracy.

**Action 8.9** *Mouse left on the Deselect All Constraints command.*

Select only the constraint we just entered.

**Action 8.10** *Mouse h-sh-left on the constraint we just entered. If there is more than one constraint in the Constraints Pane the constraint we just entered is the one at the top.*

**Action 8.11** *Mouse left on the Verify Constraints command.*

The system reports that our constraint has been verified.

**All constraints were verified.**

**Action 8.12** *Mouse left on the Transform Constraints command.*

The system coerces our constraint into a transformed constraint, adds it to the Target Transformation Section's Transformed Constraints Pane, and changes configurations to the Target Transformation Section.

**Action 8.13** *Mouse left on the Deselect All Transformed Constraints command.*

**Action 8.14** *Mouse left on the Deselect All Techniques command.*

Our constraint is at the top of the Transformed Constraints Pane.

**Action 8.15** *Mouse h-sh-left on our new constraint. It is located at the top of the Transformed Constraints Pane.*

Appropriate estimation techniques are now selected.

**Action 8.16** *Mouse h-sh-left on the  $Pc^* \rightarrow ()$  - Joback  $Pc^*$  - Group Contribution Technique.*

**Action 8.17** *Mouse h-sh-left on the  $Pc \rightarrow (Pc^*)$  - Joback  $Pc$  Modification - Equation Oriented Technique.*

**Action 8.18** *Mouse h-sh-left on the  $Tbr^* \rightarrow ()$  - Joback  $Tbr^*$  - Group Contribution Technique.*

**Action 8.19** *Mouse h-sh-left on the  $Tbr \rightarrow (Tbr^*)$  - Joback  $Tbr$  Modification - Equation Oriented Technique.*

**Action 8.20** *Mouse h-sh-left on the  $Tb \rightarrow ()$  - Joback  $Tb$  - Group Contribution Technique.*

**Action 8.21** *Mouse left on the Apply Selected Techniques command.*

**Action 8.22** *Mouse h-sh-left on the Pvp → (Tb Tbr Pc) – Riedel Plank Miller Tbr Biased – Equation Oriented Technique.*

Our constraint is now dependent only on the three fundamental physical properties:  $T_b$ ,  $T_{b^*}$ , and  $P_c^*$ .

**Action 8.23** *Mouse left on the Make Automatic command.*

The system reports that our constraint has been reduced to fundamental properties.

**All Selected Constraints have been Reduced to Fundamental Properties.**

The system then changes to the Automatic Design Section's Meta-Groups Configuration. Figure 8.9 shows the initial state of the configuration.

### Entering Meta-Molecules

The Target Transformation Section's **Make Automatic** command changes changes to the Meta-Groups Configuration. The single meta-group containing all the consistent groups is displayed in the Meta-Groups Display Pane. The initial set of meta-molecules is now entered.

We first change to the Meta-Molecules Configuration.

**Action 8.24** *Mouse right on an empty area of the screen.*

A menu is exposed displaying the system's configurations.

**Action 8.25** *Mouse left on Automatic Design Section Meta Molecules Configuration.*

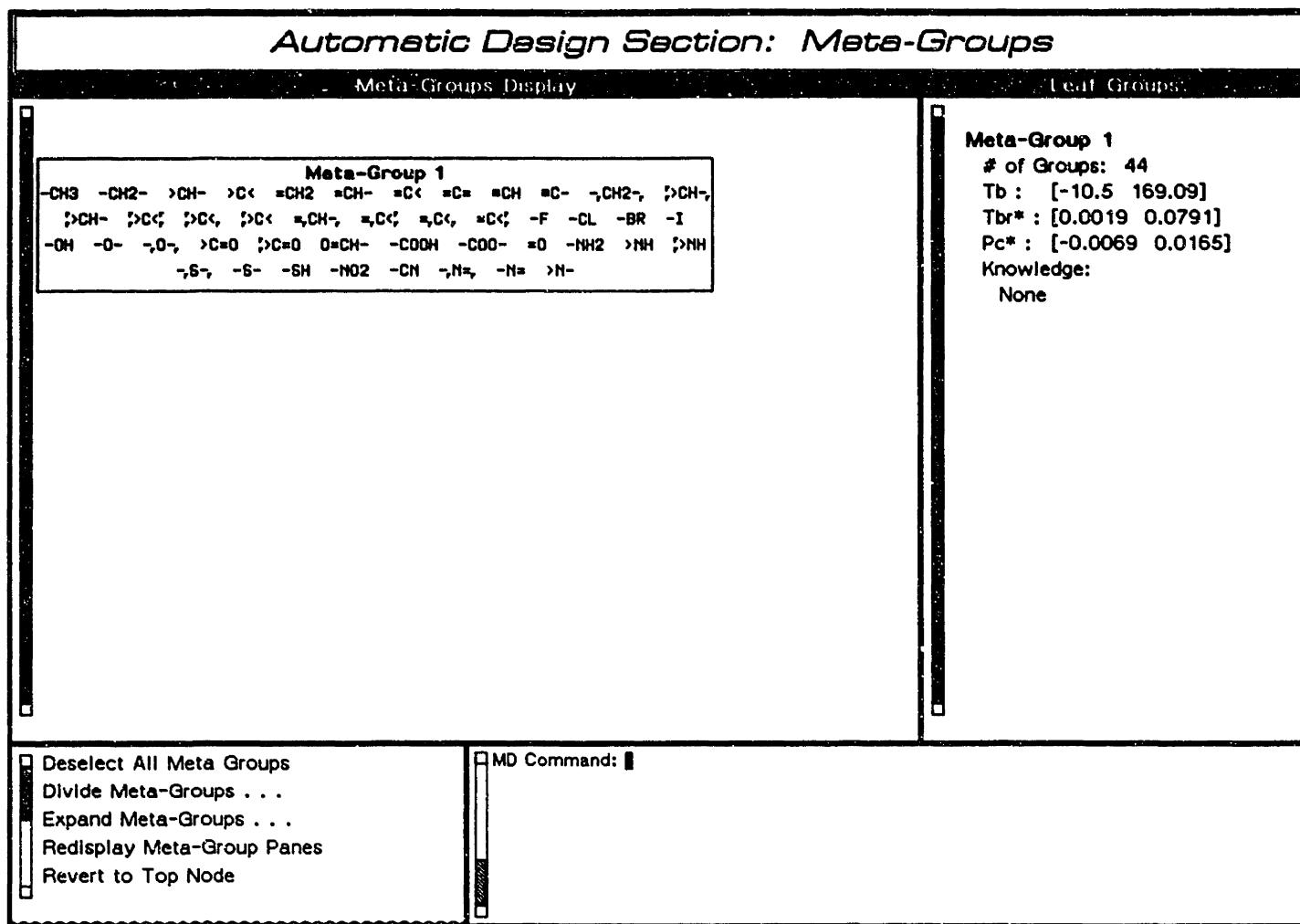


Figure 8.9: Initial Meta-Group for Automatic Design

The configuration changes to the Meta-Molecules Configuration. Our constraint is displayed in the Property Constraints Pane.

We input four meta-molecules having occurrences of our initial meta-group ranging from 2 to 5.

**Action 8.26** *Mouse left on the Input Meta-Molecules command.*

The system displays a menu prompting for the lower and upper limits on the number of occurrences.

**# of Groups Lower Bound:** *some value*

**# of Groups Upper Bound:** *some value*

**ABORT** aborts **END** uses these values.

The actual values displayed after the prompts are not important.

**Action 8.27** *Mouse left on the phrase displayed after the # of Groups Lower Bound: prompt. The phrase is replaced by a blinking cursor.*

**Action 8.28** *Enter the number 2. Press the return key.*

The process is repeated for the upper limit.

**Action 8.29** *Mouse left on the phrase displayed after the # of Groups Upper Bound: prompt. The phrase is replaced by a blinking cursor.*

**Action 8.30** *Enter the number 5. Press the return key.*

Now that both entries are complete the system creates four meta-molecules having two through five occurrences of our initial meta-group.

**Action 8.31** *Press the <END> key.*

The four meta-molecules are added to the Meta Molecules Pane. Figure 8.10 shows the four initial meta-molecules in the Meta Molecules Pane.

We return to the Meta Groups Configuration. We begin separating our initial meta-group into smaller children meta-groups.

**Action 8.32** *Type the command Previous Section into the interaction pane. Press return.*

**Action 8.33** *Mouse left on the Expand Meta Groups ... command.*

A menu is exposed displaying possible expansion strategies.

**Action 8.34** *Mouse left on Ring Class.*

Our initial meta-group is separated into three new meta-groups. Each meta-group contains groups of a consistent ring class.

Our meta-molecules have also been expanded.

**Action 8.35** *Type the command Previous Section into the interaction pane. Press return.*

Our initial set of 4 meta-molecules was expanded into 52 meta-molecules. A meta-molecule represents the number of groups which must be chosen from a particular meta-group. The meta-molecule (2) represented the set of all molecules which could be formed by choosing any two groups from meta-group 1. Now that meta-group 1 was separated into three meta-groups meta-molecule (2) is expanded into six meta-molecules:

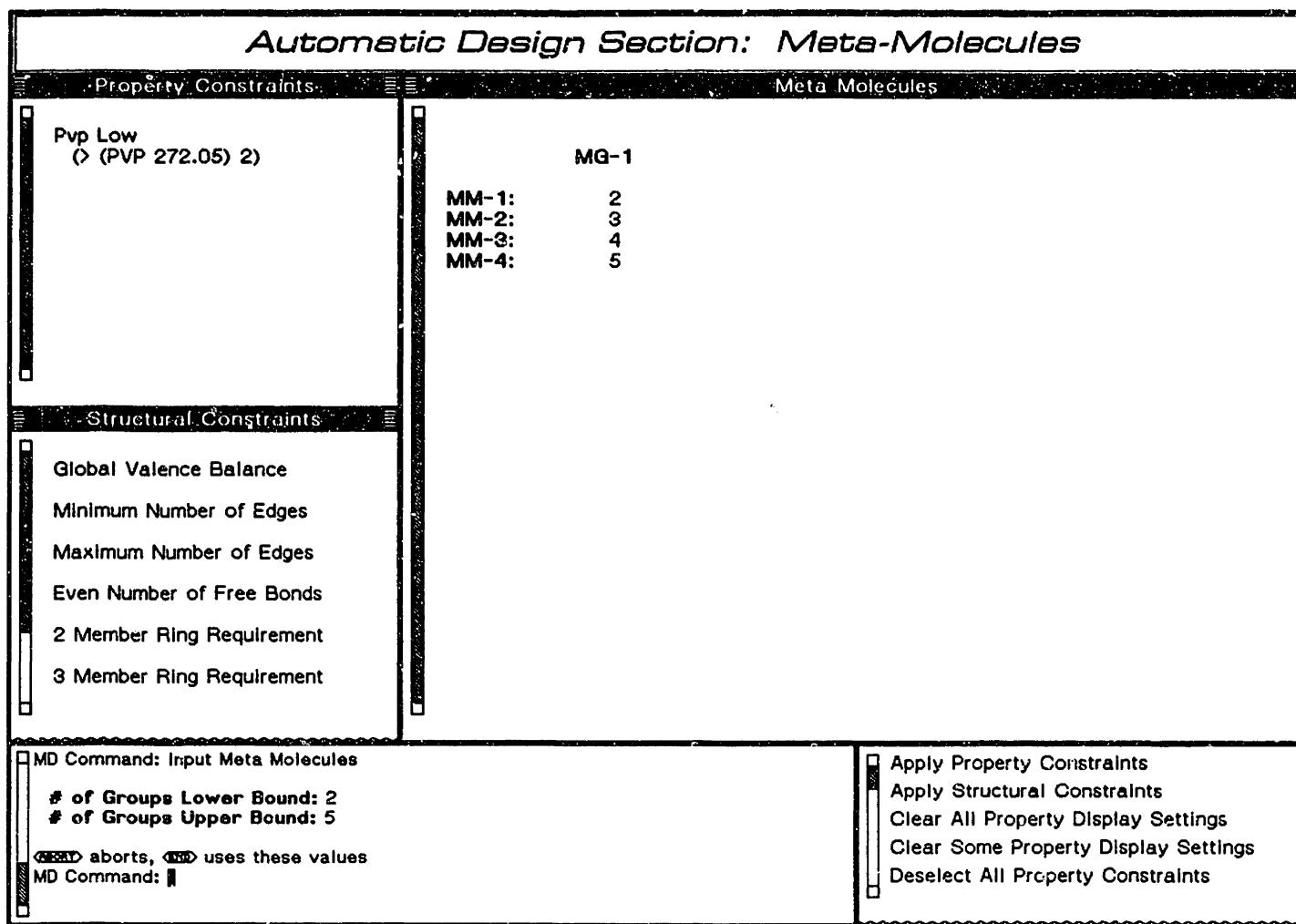


Figure 8.10: Initial Meta-Molecules for Automatic Design

(2 0 0)	(0 2 0)	(0 0 2)
(1 1 0)	(0 1 1)	(1 0 1)

## Applying Structural Constraints

Structural constraints check each meta-molecule with rules derived from graph theory.

Each constraint has a relationship between the number of nodes(groups) and the number of edges(free bonds) which must be satisfied for a meta-molecule to be structurally feasible. This rule can be examined for each structural constraint.

**Action 8.36** *Mouse left on the Document Structural Constraint command.*

The system prompts for a structural constraint.

**Enter a Structural Constraint:**

**Action 8.37** *Mouse left on the 3 Member Ring Requirement structural constraint.*

A window resource is exposed over the Meta-Molecules Display Pane. The rule used by the structural constraint is displayed. The knowledge required for the constraint to be applicable is also displayed. The 3 Member Ring Requirement structural constraint requires the ring class to be known. Having expanded by ring class we can use this constraint.

**Action 8.38** *Press the space bar.*

We now use structural constraints to prune our meta-molecules. The **Apply Structural Constraints** command uses the selected constraints of the Structural Constraints Pane.

**Action 8.39** *Mouse h-s-left on the 3 Member Ring Requirement structural constraint.*

**Action 8.40** *Mouse h-s-left on the Necessity of Mixed Groups structural constraint.*

**Action 8.41** *Mouse h-s-left on the Necessity of Non-Mixed Groups structural constraint.*

We now apply these constraints to our 52 meta-molecules.

**Action 8.42** *Mouse left on the Apply Structural Constraints command.*

A window resource is exposed over the Meta-Molecules Pane displaying the results of applying each constraint. The results are:

Constraint	Molecules		
	Checked	Kept	Pruned
3 Member Ring Requirement	52	32	20
Necessity of Mixed Groups	32	29	3
Necessity of Non-Mixed Groups	29	26	3

The Meta-Molecules Display Pane is redisplayed showing the 26 remaining meta-molecules. Figure 8.11 shows the 26 meta-molecules displayed in the Meta Molecules Pane.

**Action 8.43** *Press the space bar.*

## Automatic Design Section: Meta-Molecules

Property Constraints		Meta Molecules		
		MG-4	MG-3	MG-2
Pvp Low	Ø (PVP 272.05) 2	MM-37: 4	1	0
		MM-38: 3	2	0
		MM-41: 0	5	0
		MM-42: 4	0	1
		MM-43: 3	1	1
		MM-44: 2	2	1
		MM-47: 3	0	2
		MM-48: 2	1	2
		MM-49: 1	2	2
		MM-51: 2	0	3
		MM-52: 1	1	3
		MM-54: 1	0	4
		MM-56: 0	0	5
		MM-22: 3	1	0
		MM-25: 0	4	0
		MM-26: 3	0	1
		MM-27: 2	1	1
		MM-30: 2	0	2
		MM-31: 1	1	2
		MM-33: 1	0	3
		MG-4	MG-3	MG-2
		MM-35: 0	0	4
		MM-14: 0	3	0

MD Command: Apply Structural Constraints  
 MD Command:

Apply Property Constraints  
 Apply Structural Constraints  
 Clear All Property Display Settings  
 Clear Some Property Display Settings  
 Deselect All Property Constraints

Figure 8.11: Meta-Molecules Surviving Structural Pruning

## Applying Property Constraints

Property constraints are the second type of constraints used to prune meta-molecules.

The **Make Automatic** command coerces each of the Target Transformation Section's selected transformed-constraints into an automatic constraint and adds it to the Meta-Molecules Configuration's Property Constraints Pane. The **Apply Property Constraints** command checks each meta-molecule for satisfaction of the selected property constraints.

We prune using the our Pvp Low property constraint.

**Action 8.44** *Mouse h-sh-left on the Pvp Low property constraint.*

**Action 8.45** *Mouse left on the Apply Property Constraints command.*

The results of the pruning are displayed in a window resource exposed over the Meta-Molecules Display Pane. The results are:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	26	10	16

The Meta Molecules Pane is redisplayed showing the 10 surviving meta-molecules.

**Action 8.46** *Press the space bar.*

## Complete Design

The following actions complete the automatic design. The time required for this example is about 2 hours. The current automatic design algorithm can still be significantly improved to shorten this time.

We separate the meta-groups again.

**Action 8.47** *Type the command Previous Section into the interaction pane. Press return.*

**Action 8.48** *Mouse left on the Expand Meta Groups ... command.*

A menu is exposed displaying expansion strategies.

**Action 8.49** *Mouse left on Global Valence.*

The meta-groups are expanded into 9 new meta-groups. We return to the Meta-Molecules Configuration for pruning.

**Action 8.50** *Type the command Previous Section into the interaction pane. Press return.*

Our 10 meta-molecules were expanded into 471 new meta-molecules. Whenever the number of meta-molecules exceeds 250 the system queries the designer if he or she really wants all the meta-molecules displayed.

**There are 471 meta-molecules. Do you wish them displayed?**

**Action 8.51** *Type No. Press return.*

The expansion by Global Valence added knowledge to the system. Additional structural constraints are now applicable.

**Action 8.52** *Mouse left on the Deselect all Structural Constraints command.*

**Action 8.53** *Mouse h-sh-left on the Global Valence Balance structural constraint.*

**Action 8.54** *Mouse h-sh-left on the Minimum Number of Edges structural constraint.*

**Action 8.55** *Mouse h-sh-left on the Maximum Number of Edges structural constraint.*

**Action 8.56** *Mouse h-sh-left on the Even Number of Free Bonds structural constraint.*

We now use these selected structural constraints to prune our meta-molecules.

**Action 8.57** *Mouse left on the Apply Structural Constraints command.*

The results of the pruning are:

Constraint	Molecules		
	Checked	Kept	Pruned
Global Valence Balance	471	357	114
Minimum Number of Edges	357	357	0
Maximum Number of Edges	357	310	47
Even Number of Free Bonds	310	165	145

The Meta-Molecules Display Pane is redisplayed.

**Action 8.58** *Press the space bar.*

We apply the Pvp Low property constraint again.

**Action 8.59** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	165	32	133

The Meta-Molecules Display Pane is redisplayed.

**Action 8.60** *Press the space bar.*

We return to the Meta-Groups Configuration to further expand meta-groups.

**Action 8.61** *Type the command Previous Section into the interaction pane. Press return.*

Meta-group 7 contains a large number of groups and has a large meta-contribution for  $T_b$ . The **Show Meta Contributions** command allows us to examine its meta-contributions.

**Action 8.62** *Mouse left on the Show Meta Contributions command.*

The system prompts for a meta-group:

**Enter a Meta-Group:**

**Action 8.63** *Mouse left on Meta Group 7 displayed in the Leaf Groups Pane.*

The group contributions for each fundamental physical property is displayed in a window resource exposed over the Meta-Groups Display Pane. We see that the  $=O$  group has a large negative contribution toward  $T_b$ .  $=O$  is also a group we do not want many occurrences of. We concentrate on isolating  $=O$ .

**Action 8.64** *Press the space bar.*

**Action 8.65** *Mouse h-sh-left on Meta-Group 7 displayed in the Leaf Groups Pane.*

**Action 8.66** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu displaying division strategies.

**Action 8.67** *Mouse left on Divide Manually.*

The system displays the meta-contributions for Meta-Group 7 and prompts for a subset of 15 groups it contains.

**Action 8.68** *Mouse left on the =O group displayed in the interaction pane.*

The =O group becomes highlighted.

**Action 8.69** *Press the <END> key.*

Meta-Group 7 is divided into two new meta-groups: Meta-Group 14 and Meta-Group 15. Meta-Group 14 contains the single group =O.

We apply our Pvp Low constraint again.

**Action 8.70** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 88 meta-molecules.

**Action 8.71** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	88	70	18

The Meta-Molecules Display Pane is redisplayed.

**Action 8.72** *Press the space bar.*

We return to the Meta-Groups Configuration to further expand meta-groups.

**Action 8.73** *Type the command Previous Section into the interaction pane. Press return.*

Meta-Group 15 still has a considerably large  $T_b$  contribution.

**Action 8.74** *Mouse h-sh-left on Meta-Group 15 displayed in the Leaf Groups Pane.*

**Action 8.75** *Mouse left on the Divide Meta-Groups command.*

The system exposes a menu displaying division strategies.

**Action 8.76** *Mouse left on Divide by Largest Gap wrt Property.*

The system prompts for a property:

**Enter the Property:**

**Action 8.77** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 15.

**Action 8.78** *Mouse left on Normal Boiling Point.*

Meta-Group 15 is divided into two new meta-groups: Meta-Group 16 and Meta-Group 17. Meta-Group 17 contains three groups,  $-\text{COOH}$ ,  $-\text{NO}_2$ , and  $-\text{CN}$ , which have relatively large contributions toward  $T_b$ . Isolating these groups should help our pruning.

We apply our Pvp Low constraint again.

**Action 8.79** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 133 meta-molecules.

**Action 8.80** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Constraint	Molecules		
	Checked	Kept	Pruned
Pvp Low	133	70	63

The Meta-Molecules Display Pane is redisplayed.

**Action 8.81** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Group 17 does not occur in any of the meta-molecules.

**Action 8.82** *Mouse left on the Show Occurrence Statistics command.*

Meta-Group 17's average occurrence is 0.0 indicating that it is absent from all meta-molecules. This is very useful information because Meta-Group 17 still contains three

groups and we possibly could have considered expanding it at some point. We now see that further expansion is unnecessary.

Meta-Group 16 is seen to have a relatively large number of groups, 11, and a relatively high average occurrence, 0.9. We examine further expansion of Meta-Group 16.

**Action 8.83** *Press the space bar.*

**Action 8.84** *Type the command Previous Section into the interaction pane.*

Meta-Group 16 has a large  $T_b$  meta-contribution.

**Action 8.85** *Mouse h-sh-left on Meta-Group 16 displayed in the Leaf Groups Pane.*

**Action 8.86** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu displaying division strategies.

**Action 8.87** *Mouse left on Divide in Half wrt Property Contributions.*

The system prompts for a property:

**Enter a Property:**

**Action 8.88** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 16.

**Action 8.89** *Mouse left on Normal Boiling Point.*

Meta-Group 16 is divided into two new meta-groups: Meta-Group 18 and Meta-Group 18. Meta-Group 18 contains five groups. Meta-Group 19 contains six groups.

We apply our Pvp Low constraint again.

**Action 8.90** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 133 meta-molecules.

**Action 8.91** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Constraint	Molecules		
	Checked	Kept	Pruned
Pvp Low	133	64	69

The Meta-Molecules Display Pane is redisplayed.

**Action 8.92** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Group 18 has a fairly large number of groups, 5, and a high average occurrence, 0.703125.

**Action 8.93** *Mouse left on the Show Occurrence Statistics command.*

We return to the Meta-Groups Configuration to expand Meta-Group 18.

**Action 8.94** *Press the space bar.*

**Action 8.95** *Type the command Previous Section into the interaction pane. Press return.*

We first examine Meta-Group 18's meta-contributions.

**Action 8.96** *Mouse left on the Show Meta-Contributions command.*

The system prompts for a meta-group.

**Enter a Meta-Group:**

**Action 8.97** *Mouse left on Meta-Group 18 displayed in the Leaf Groups Pane.*

The meta-contributions are displayed on a window resource exposed over the Meta-Groups Display Pane.

Examining the meta-contribution for  $T_b$  we see that the largest gap is between  $-\text{CH}_3$  and  $-\text{Cl}$ . However, the second largest gap, between  $-\text{F}$  and  $\equiv\text{CH}$ , also separates the meta-contribution in positive and negative values. We divide Meta-Group 18 by the sign of contributions.

**Action 8.98** *Press the space bar.*

**Action 8.99** *Mouse h-sh-left on Meta-Group 18 displayed in the Leaf Groups Pane.*

**Action 8.100** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu displaying division strategies.

**Action 8.101** *Mouse left on Divide by Sign of Contributions wrt Property.*

The system prompts for a property:

**Enter a Property:**

**Action 8.102** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 18.

**Action 8.103** *Mouse left on Normal Boiling Point.*

Meta-Group 18 is divided into two new meta-groups: Meta-Group 20 and Meta-Group 21. Meta-Group 21 contains the single group:  $-F$ .

We apply our Pvp Low constraint again.

**Action 8.104** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 109 meta-molecules.

**Action 8.105** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	109	93	16

The Meta-Molecules Display Pane is redisplayed.

**Action 8.106** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Group 8 has a large number of groups, 10, and a high average occurrence, 0.333333.

**Action 8.107** *Mouse left on the Show Occurrence Statistics command.*

We return to the Meta-Groups Configuration to expand Meta-Group 8.

**Action 8.108** *Press the space bar.*

**Action 8.109** *Type the command Previous Section into the interaction pane. Press return.*

We first examine Meta-Group 8's meta-contributions.

**Action 8.110** *Mouse left on the Show Meta-Contributions command.*

The system prompts for a meta-group.

**Enter a Meta-Group:**

**Action 8.111** *Mouse left on Meta-Group 8 displayed in the Leaf Groups Pane.*

The meta-contributions are displayed on a window resource exposed over the Meta-Groups Display Pane. The contributions toward  $T_b$  seem to have a large gap between 27.38 and 50.17.

**Action 8.112** *Press the space bar.*

**Action 8.113** *Mouse h-sh-left on Meta-Group 8 displayed in the Leaf Groups Pane.*

**Action 8.114** *Mouse left on the Divide Meta-Groups command.*

The system exposes a menu displaying division strategies.

**Action 8.115** *Mouse left on Divide by Largest Gap wrt Property.*

The system prompts for a property:

**Enter the Property:**

**Action 8.116** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 8.

**Action 8.117** *Mouse left on Normal Boiling Point.*

Meta-Group 8 is divided into two new meta-groups: Meta-Group 22 and Meta-Group 23.

We apply our Pvp Low constraint again.

**Action 8.118** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 124 meta-molecules.

**Action 8.119** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Constraint	Molecules		
	Checked	Kept	Pruned
Pvp Low	124	104	20

The Meta-Molecules Display Pane is redisplayed.

**Action 8.120** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Group 14 has a very high average occurrence, 1.2211539. Yet it only has one group. In fact Meta-Group 14 contains only the =O group.

**Action 8.121** *Mouse left on the Show Occurrence Statistics command.*

To confirm this we briefly return to the Meta-Groups Configuration.

**Action 8.122** *Press the space bar.*

**Action 8.123** *Type the command Previous Section into the interaction pane.*

The Meta-Groups Display Pane shows that Meta-Group 14 contains only =O.

**Action 8.124** *Type the command Previous Section into the interaction pane.*

Examining the occurrences in the Meta Molecules Pane we see that Meta-Group 14 occurs sometimes as many as four times in a single meta-molecule. This is a chemically unlikely situation. We limit the occurrence of =O to a maximum of 1.

**Action 8.125** *Mouse left on the Limit Group Occurrences command.*

The system exposes a menu displaying prompts for each of the current meta-groups.

The default value for each of the limits is `nil` indicating no limit.

**Action 8.126** *Mouse left on the `nil` following the MG-14: prompt. The `nil` is replaced by a blinking cursor.*

**Action 8.127** *Type 1. Press return.*

**Action 8.128** *Press the <END> key.*

All meta-molecules which contained more than 1 occurrence of Meta-Group 14 are removed from the Meta Molecules Pane. There are now 67 meta-molecules.

We return to the Meta-Groups Configuration.

**Action 8.129** *Type the command Previous Section into the interaction pane. Press return.*

Meta-Group 19 has large meta-contributions to both  $T_b$  and  $T_{b_r}^*$ . We divide it in half.

**Action 8.130** *Mouse h-sh-left on Meta-Group 19 displayed in the Leaf Groups Pane.*

**Action 8.131** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu displaying division strategies.

**Action 8.132** *Mouse left on Divide in Half wrt Property Contributions.*

The system prompts for a property:

**Enter a Property:**

**Action 8.133** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 19.

**Action 8.134** *Mouse left on Joback Modified Reduced Boiling Point.*

Meta-Group 19 is divided into two new meta-groups: Meta-Group 24 and Meta-Group 25. Meta-Group 25 contains the single group:  $-\text{OH}$ .

Examining the meta-group statistics we see that the total number of meta-molecules increased only by 1.

**Action 8.135** *Mouse left on the Show Meta-Group Statistics command.*

The system exposes a window resource over the Meta-Groups Display Pane displaying several statistics. The total number of meta-molecules is 68.

**Action 8.136** *Press the space bar.*

We divide Meta-Group 24.

**Action 8.137** *Mouse h-sh-left on Meta-Group 24 displayed in the Leaf Groups Pane.*

**Action 8.138** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu displaying division strategies.

**Action 8.139** *Mouse left on Divide in Half wrt Property Contributions.*

The system prompts for a property:

**Enter a Property:**

**Action 8.140** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 24.

**Action 8.141** *Mouse left on Normal Boiling Point.*

Meta-Group 24 is divided into two new meta-groups: Meta-Group 26 and Meta-Group 27.

We continue to expand meta-groups.

**Action 8.142** *Mouse h-sh-left on Meta-Group 20 displayed in the Leaf Groups Pane.*

**Action 8.143** *Mouse left on the Divide Meta-Groups command.*

The system exposes a menu displaying division strategies.

**Action 8.144** *Mouse left on Divide by Largest Gap wrt Property.*

The system prompts for a property:

**Enter the Property:**

**Action 8.145** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Group 20.

**Action 8.146** *Mouse left on Normal Boiling Point.*

Meta-Group 20 is divided into two new meta-groups: Meta-Group 28 and Meta-Group 29.

We apply our Pvp Low constraint again.

**Action 8.147** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 113 meta-molecules.

**Action 8.148** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

## Molecules

Constraint	Checked	Kept	Pruned
Pvp Low	113	76	37

The Meta-Molecules Display Pane is redisplayed.

**Action 8.149** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Groups 27, 25, and 17 all have zero average occurrences.

**Action 8.150** *Mouse left on the Show Occurrences Statistics command.*

Meta-Group 22 has a high average occurrence of 0.3026316. Meta-Group 11 has a lower average occurrence, 0.09210526, but a relatively large number of groups. We pursue the expansion of Meta-Groups 22 and 11.

**Action 8.151** *Press the space bar.*

**Action 8.152** *Type the command Previous Section into the interaction pane.*

**Action 8.153** *Mouse h-sh-left on Meta-Group 22 displayed in the Leaf Groups Pane.*

**Action 8.154** *Mouse h-sh-left on Meta-Group 11 displayed in the Leaf Groups Pane.*

**Action 8.155** *Mouse left on the Divide Meta-Groups command.*

The system exposes a menu displaying division strategies.

**Action 8.156** *Mouse left on Divide by Largest Gap wrt Property.*

The system prompts for a property:

**Enter the Property:**

**Action 8.157** *Mouse right on a blank area within the interaction pane.*

The system exposes a menu of the fundamental properties contained in Meta-Groups 22 and 11.

**Action 8.158** *Mouse left on Normal Boiling Point.*

Meta-Group 22 is divided into two new meta-groups: Meta-Group 30 and Meta-Group 31. Meta-Group 11 is divided into two new meta-groups: Meta-Group 32 and Meta-Group 33.

Meta-Groups 28, 26, and 32 all contain a relatively large number of groups.

**Action 8.159** *Mouse h-sh-left on Meta-Group 28 displayed in the Leaf Groups Pane.*

**Action 8.160** *Mouse h-sh-left on Meta-Group 26 displayed in the Leaf Groups Pane.*

**Action 8.161** *Mouse h-sh-left on Meta-Group 32 displayed in the Leaf Groups Pane.*

**Action 8.162** *Mouse left on the Divide Meta-Groups command.*

The system exposes a menu displaying division strategies.

**Action 8.163** *Mouse left on Divide in Half.*

We apply our Pvp Low constraint again.

**Action 8.164** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 175 meta-molecules.

**Action 8.165** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Constraint	Molecules		
	Checked	Kept	Pruned
Pvp Low	175	114	61

The Meta-Molecules Display Pane is redisplayed.

**Action 8.166** *Press the space bar.*

Examining the occurrence statistics shows that Meta-Groups 37, 27, 25, 17, 38, 39, and 33 all have zero average occurrences.

**Action 8.167** *Mouse left on the Show Occurrences Statistics command.*

Meta-Groups 14, 35, 29, 21, 10, and 13 all contain a single group. We continue to expand the remaining groups.

**Action 8.168** *Press the space bar.*

**Action 8.169** *Type the command Previous Section into the interaction pane. Press return.*

**Action 8.170** *Mouse h-sh-left on Meta-Group 5 displayed in the Leaf Groups Pane.*

**Action 8.171** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.172** *Mouse left on Divide in Half.*

**Action 8.173** *Mouse h-sh-left on Meta-Group 40 displayed in the Leaf Groups Pane.*

**Action 8.174** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.175** *Mouse left on Divide in Half.*

**Action 8.176** *Mouse left on the Show Meta-Group Statistics command.*

The total number of Meta-Molecules is 150.

We expand two more meta-groups.

**Action 8.177** *Press the space bar.*

**Action 8.178** *Mouse h-sh-left on Meta-Group 6 displayed in the Leaf Groups Pane.*

**Action 8.179** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.180** *Mouse left on Divide in Half.*

**Action 8.181** *Mouse left on the Show Meta-Group Statistics command.*

The total number of Meta-Molecules is now 178.

**Action 8.182** *Press the space bar.*

**Action 8.183** *Mouse h-sh-left on Meta-Group 34 displayed in the Leaf Groups Pane.*

**Action 8.184** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.185** *Mouse left on Divide in Half.*

**Action 8.186** *Mouse left on the Show Meta-Group Statistics command.*

The total number of Meta-Molecules is now 249.

**Action 8.187** *Press the space bar.*

We apply our Pvp Low constraint again.

**Action 8.188** *Type the command Previous Section into the interaction pane.*

The Meta Molecules Pane shows 249 meta-molecules.

**Action 8.189** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Constraint	Molecules		
	Checked	Kept	Pruned
Pvp Low	249	168	81

The Meta-Molecules Display Pane is redisplayed.

**Action 8.190** *Press the space bar.*

Six meta-groups remain with non-zero occurrences and multiple groups: 36, 30, 31, 23, 9, and 12.

**Action 8.191** *Mouse left on the Show Occurrences Statistics command.*

These are expanded.

**Action 8.192** *Press the space bar.*

**Action 8.193** *Type the command Previous Section into the interaction pane. Press return.*

**Action 8.194** *Mouse h-sh-left on Meta-Group 36 displayed in the Leaf Groups Pane.*

**Action 8.195** *Mouse h-sh-left on Meta-Group 30 displayed in the Leaf Groups Pane.*

**Action 8.196** *Mouse h-sh-left on Meta-Group 31 displayed in the Leaf Groups Pane.*

**Action 8.197** *Mouse h-sh-left on Meta-Group 23 displayed in the Leaf Groups Pane.*

**Action 8.198** *Mouse h-sh-left on Meta-Group 9 displayed in the Leaf Groups Pane.*

**Action 8.199** *Mouse h-sh-left on Meta-Group 12 displayed in the Leaf Groups Pane.*

**Action 8.200** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.201** *Mouse left on Divide in Half.*

**Action 8.202** *Mouse left on the Show Meta-Group Statistics command.*

The total number of Meta-Molecules is 349.

**Action 8.203** *Press the space bar.*

We apply our Pvp Low constraint again.

**Action 8.204** *Type the command Previous Section into the interaction pane.*

Since the number of meta-molecules is greater than 250 the system queries for confirmation of display.

**There are 349 meta-molecules. Do you wish them displayed?**

**Action 8.205** *Type No. Press return.*

**Action 8.206** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	349	279	70

The Meta-Molecules Display Pane is redisplayed.

**Action 8.207** *Press the space bar.*

The system queries for confirmation of display.

**There are 279 meta-molecules. Do you wish them displayed?**

**Action 8.208** *Type No. Press return.*

Three meta-groups are left needing expansion: 52, 54, and 56. These are expanded.

**Action 8.209** *Type the command Previous Section into the interaction pane. Press return.*

**Action 8.210** *Mouse h-sh-left on Meta-Group 52 displayed in the Leaf Groups Pane.*

**Action 8.211** *Mouse h-sh-left on Meta-Group 54 displayed in the Leaf Groups Pane.*

**Action 8.212** *Mouse h-sh-left on Meta-Group 56 displayed in the Leaf Groups Pane.*

**Action 8.213** *Mouse left on the Divide Meta-Groups ... command.*

The system exposes a menu of division strategies.

**Action 8.214** *Mouse left on Divide in Half.*

We apply our Pvp Low constraint for the last time.

**Action 8.215** *Type the command Previous Section into the interaction pane.*

The system queries for confirmation of display.

**There are 379 meta-molecules. Do you wish them displayed?**

**Action 8.216** *Type No. Press return.*

**Action 8.217** *Mouse left on the Apply Property Constraints command.*

The result of the pruning is:

Molecules			
Constraint	Checked	Kept	Pruned
Pvp Low	379	357	22

The Meta-Molecules Display Pane is redisplayed.

**Action 8.218** *Press the space bar.*

The system queries for confirmation of display.

**There are 357 meta-molecules. Do you wish them displayed?**

**Action 8.219** *Type No. Press return.*

Now that all the meta-groups have been fully expanded we can generate our final report.

**Action 8.220** *Mouse left on the Generate Final Report command.*

The system prompts for a file in which to store the report.

Table 8.1: 20 Automatically Designed Molecules

Meta-Molecule	Occurrence(Group)
MM-3079	1(=O) 1(-F)
MM-3078	1(=O) 1(-CL)
MM-3077	1(=O) 1(-CH3)
MM-3076	1(=O) 1(=CH2)
MM-3075	1(=O) 1( $\equiv$ CH)
MM-3074	1(=O) 1(-SH)
MM-3073	1(=O) 1(-BR)
MM-3072	2(-F)
MM-3071	1(-CL) 1(-F)
MM-3070	1(-CH3) 1(-F)
MM-3069	1(=CH2) 1(-F)
MM-3068	1( $\equiv$ CH) 1(-F)
MM-3067	1( $\equiv$ CH) 1(-CL)
MM-3066	2(-CH3)
MM-3065	1(=CH2) 1(-CH3)
MM-3064	1( $\equiv$ CH) 1(-CH3)
MM-3063	2(=CH2)
MM-3062	1( $\equiv$ CH) 1(=CH2)
MM-3061	2( $\equiv$ CH)
MM-3060	1(=O) 1(-F) 1(-O-)

Table 8.2: Seven Surviving Automatically Designed Molecules

Meta-Molecule	Occurrence(Group)	
MM-3076	1(=O)	1(=CH2)
MM-3072	2(-F)	
MM-3071	1(-CL)	1(-F)
MM-3070	1(-CH3)	1(-F)
MM-3066	2(-CH3)	
MM-3063	2(=CH2)	
MM-3061	2( $\equiv$ CH)	

**Enter the file for storage:**

**Action 8.221** *Type Fungus:>kevin>Molecule-Report.lisp. Press return.*

The first 20 meta-molecules contained in the report are shown in Table 8.1. It is apparent from Table 8.1 that the structural constraints are not complete. The information which is lacking is that of bond compatibility. Additional pruning must be done by the designer. Figure 8.2 shows the seven molecules remaining after pruning.

Figure 8.12 shows a portion of the final meta-groups tree developed during the automatic design.

## Automatic Design Section: Meta-Groups

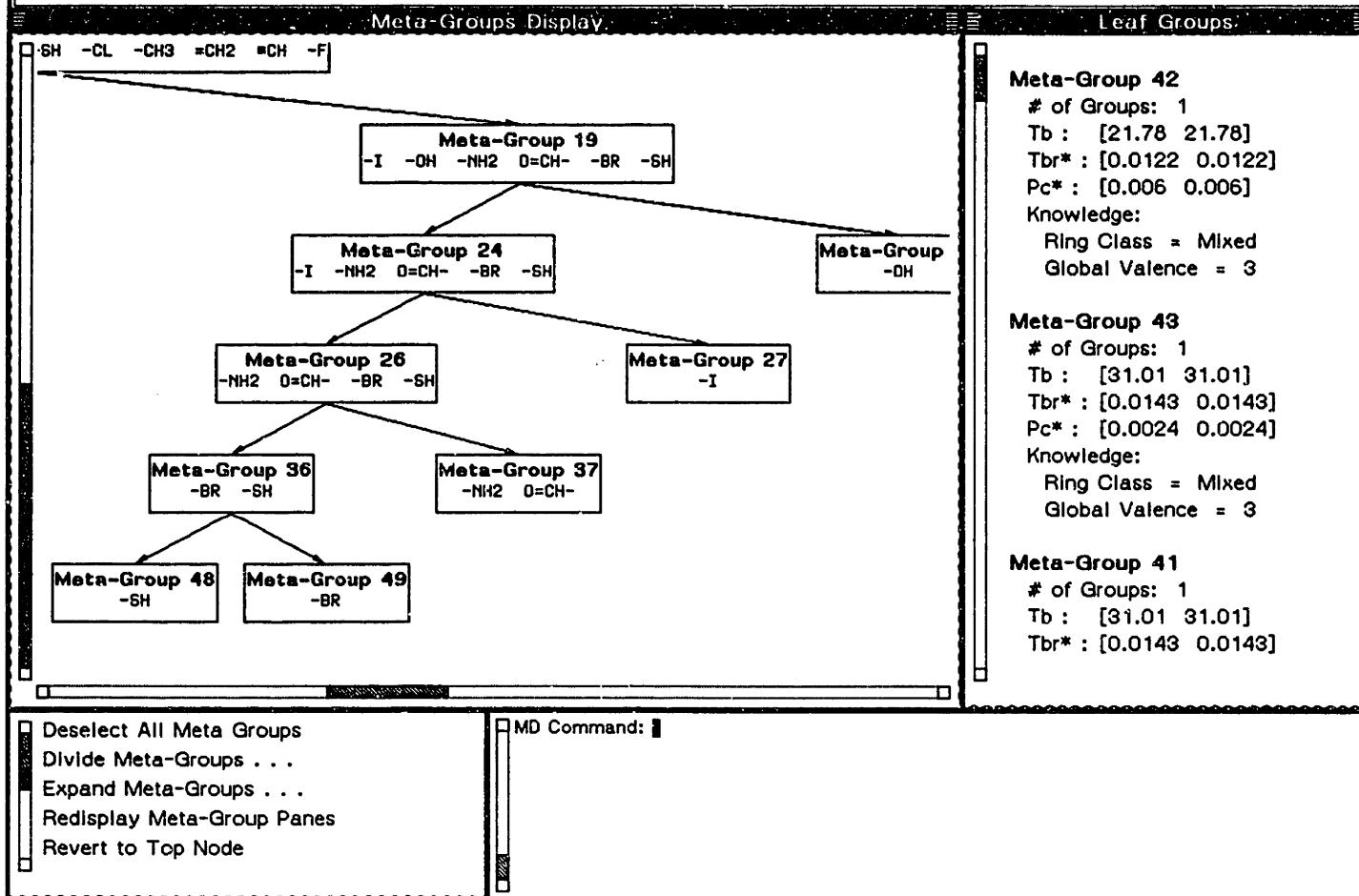


Figure 8.12: Final Display of Meta-Groups Tree after Automatic Design

# Chapter 9

## Interactive Design Section

Interactive design is performed in a fundamental physical property space. A fundamental physical property space is a graph in which each of the axes corresponds to a fundamental physical property. In such a space a group's contributions form a vector. After suitable target transformation design constraints can be displayed in this space. The objective of interactive design is to select a collection of group vectors whose vector sum satisfy these design constraints.

The interactive design methodology requires displaying and manipulating physical property spaces and group vectors. The two configurations of the Interactive Design Section address these tasks. The Preparation Configuration accepts the initial specifications for physical property design spaces. These specifications include constraints to be displayed, the grid size of the solution procedure, and upper and lower bounds for physical property values. The Design Configuration displays the design spaces and group vectors. The configuration provides facilities to assist in selecting groups and in manipulating design constraints. I describe both of these configurations.

## 9.1 Section Layout

The screen layout of the Preparation Configuration is shown in Figure 9.1. The screen real estate is used by five panes:

**Interactive Design Section Preparation Configuration Title Pane:** Displays the title of the Preparation Configuration.

**Interactive Constraints Pane:** Displays a list of interactive constraints known to the system. The **Make Interactive** command of the Target Transformation Section adds transformed constraints into this pane.

**Design Pane Specifications Pane:** Displays design specification objects. These objects are used to create physical property design panes.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Interactive Design Section Preparation Configuration Commands Pane:** The command menu displaying commands relevant to the Preparation Configuration.

The screen layout of the Design Configuration is shown in Figure 9.2. The screen real estate is used by five panes:

**Interactive Design Section Design Configuration Title Pane:** Displays the title of the Design Configuration.

## *Interactive Design Section: Preparation Configuration*

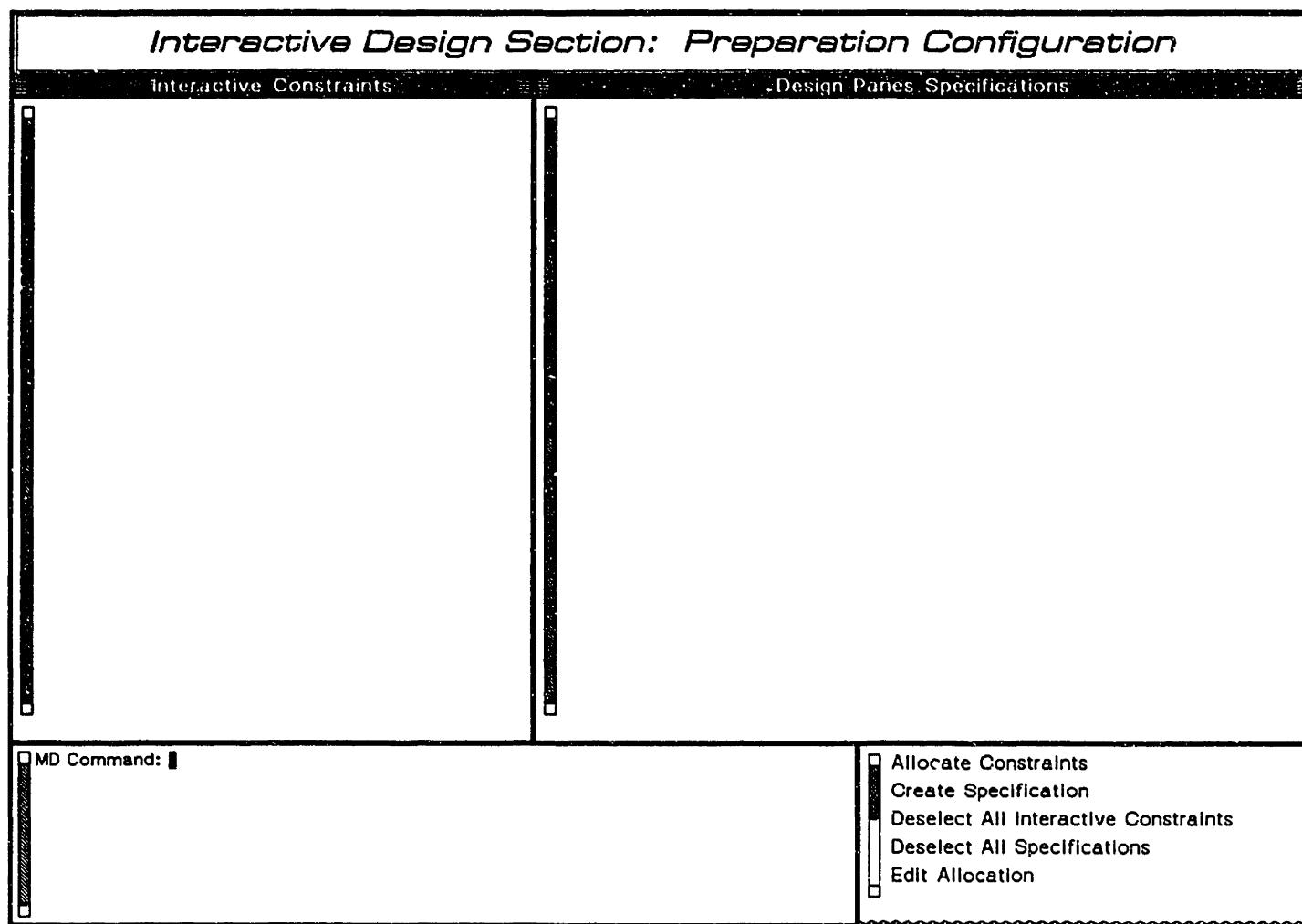


Figure 9.1: Interactive Design Section Preparation Configuration Screen

## Interactive Design Section: Design Configuration

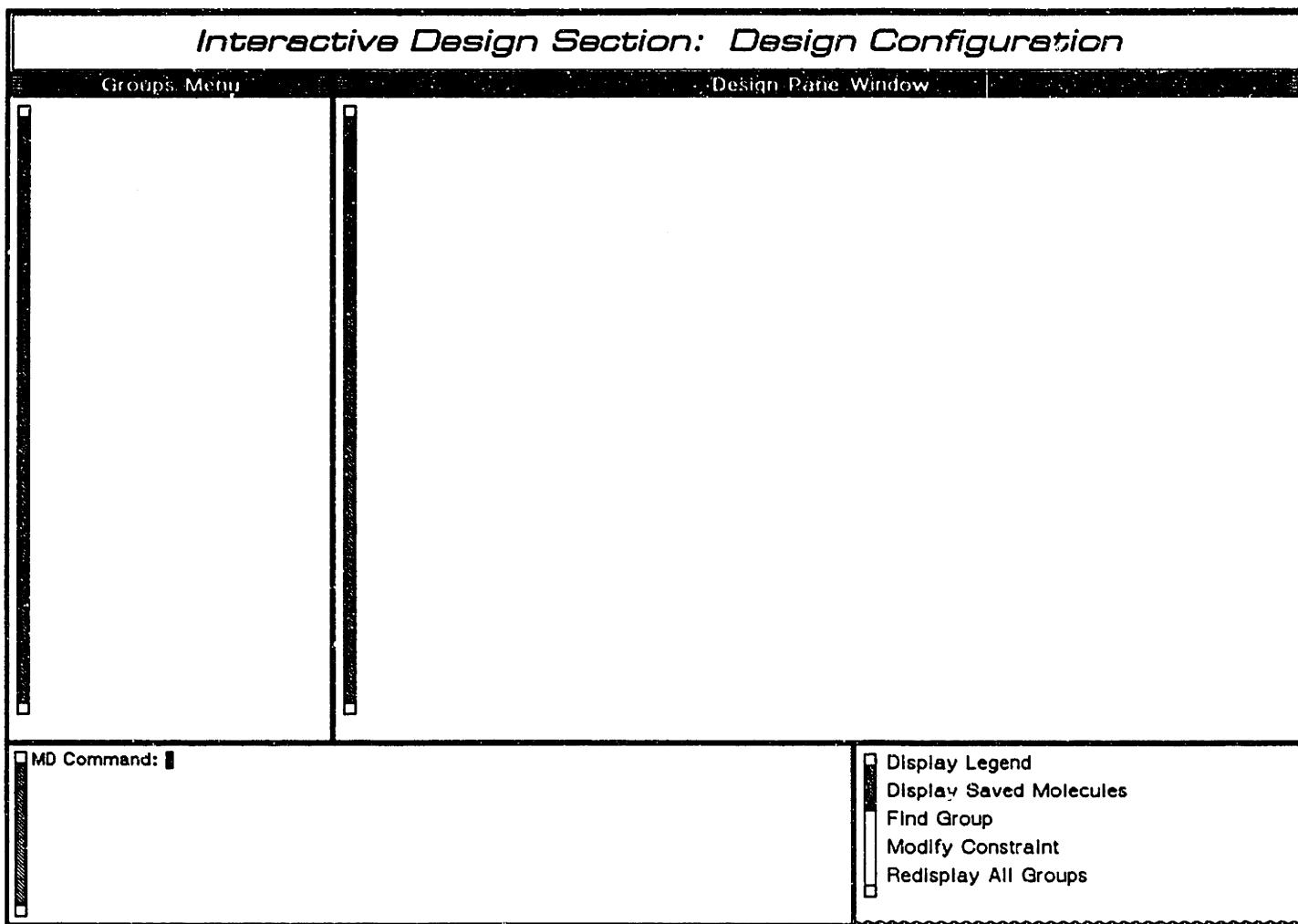


Figure 9.2: Interactive Design Section Design Configuration Screen

**Groups Menu Pane:** Displays the set of groups consistent with the design spaces displayed in the Design Pane Window Pane. When the designer moves the mouse over a group displayed in this pane a temporary group vector is displayed in every design space. An after daemon on the pane's `:mouse-moves` method implements this display.

Whenever the mouse moves in the Groups Menu Pane a `:mouse-moves` message is sent to the pane. The after daemon checks if a group is presented at the current mouse location. This is done using Symbolics' function `:displayed-presentation-at-position`. If a presentation is under the mouse the group object is extracted and displayed in all the design spaces.

The groups displayed in the Groups Menu Pane are manipulated by a variety of commands. These commands either change the order in which groups are presented or remove groups from the pane.

**Design Pane Window Pane:** Displays physical property design spaces. Commands are provided to display the legends, move, and resize the design spaces displayed in this pane.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Interactive Design Section Design Configuration Commands Pane:** The command menu displaying commands relevant to the Design Configuration.

## 9.2 Section Operation

The target transformation section's **Make Interactive** command coerces **transformed-constraints** into interactive constraints, adds these to the Preparation Configuration's Constraints Pane, and changes to the Preparation Configuration. The Preparation Configuration is used to specify information about the design spaces to be used in interactive design. This information is stored in a **design-pane-specification-object**. The **Create Specification** command creates a new specification and adds it to the Design Panes Specifications Pane. Figure 9.3 shows the Preparation Configuration with four interactive constraints and a newly created specifications object.

The designer chooses which interactive constraints are to be presented together in a design pane. This choice is done by allocating interactive constraints to a **design-pane-specification-object**. The **Allocate Constraints** command interfaces this choice. Figure 9.4 shows our four interactive constraint allocated to our specification object. Once the desired constraints have been allocated the **Form Design Panes** command extracts the information from the specification object, creates a design pane, adds this pane to the Design Configuration's Design Pane Window Pane, and changes to the Design Configuration.

The feasible region for an interactive constraint is solved for using a grid solution procedure. The upper and lower limits on the fundamental physical properties used for axes determine the solution space. This space is discretized in both the horizontal and vertical directions. Figure 9.5 shows an example discretization.

The fundamental physical properties which compose the space are  $F_1$  and  $F_2$ . The

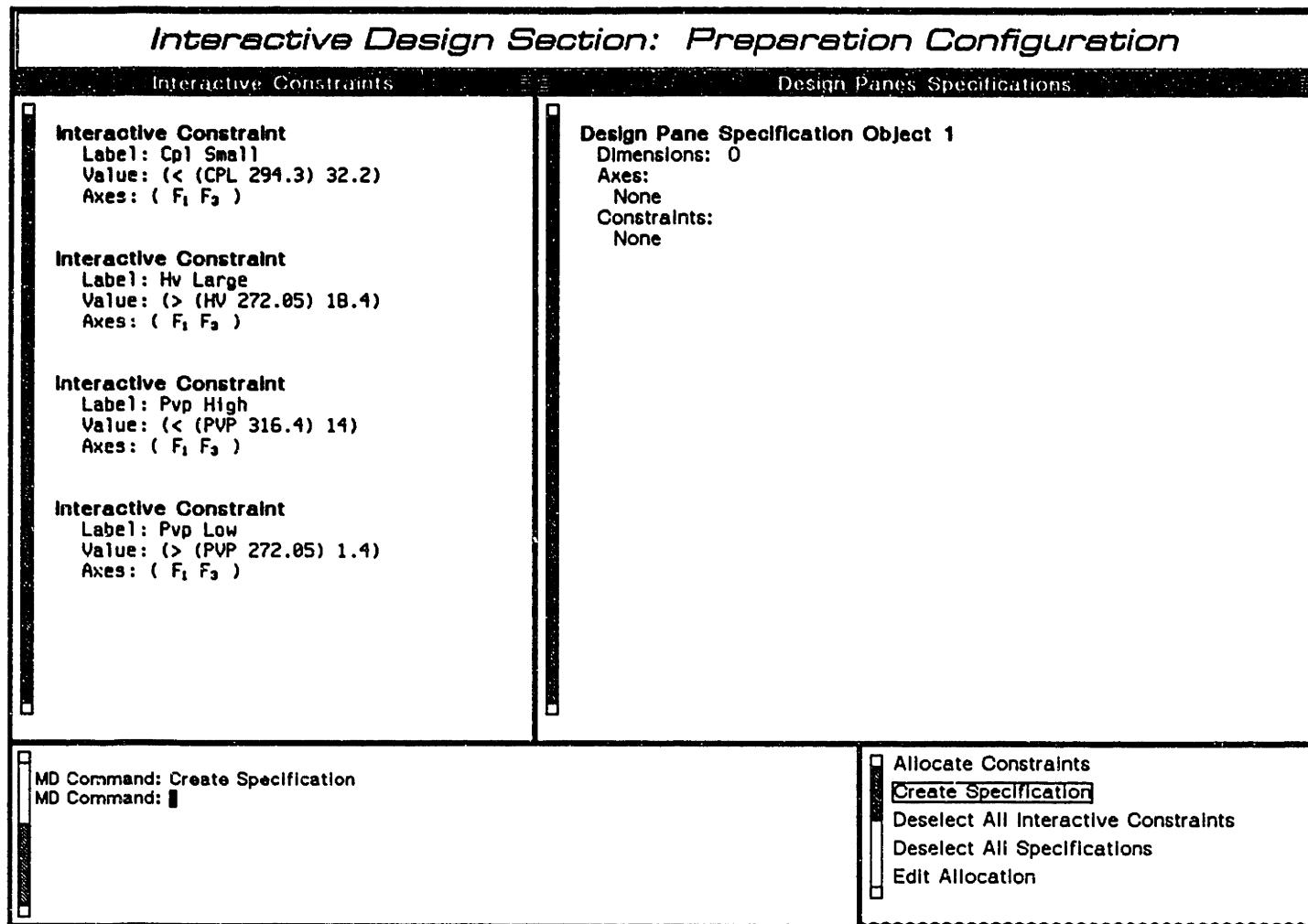


Figure 9.3: Interactive Refrigerant Design Preparation

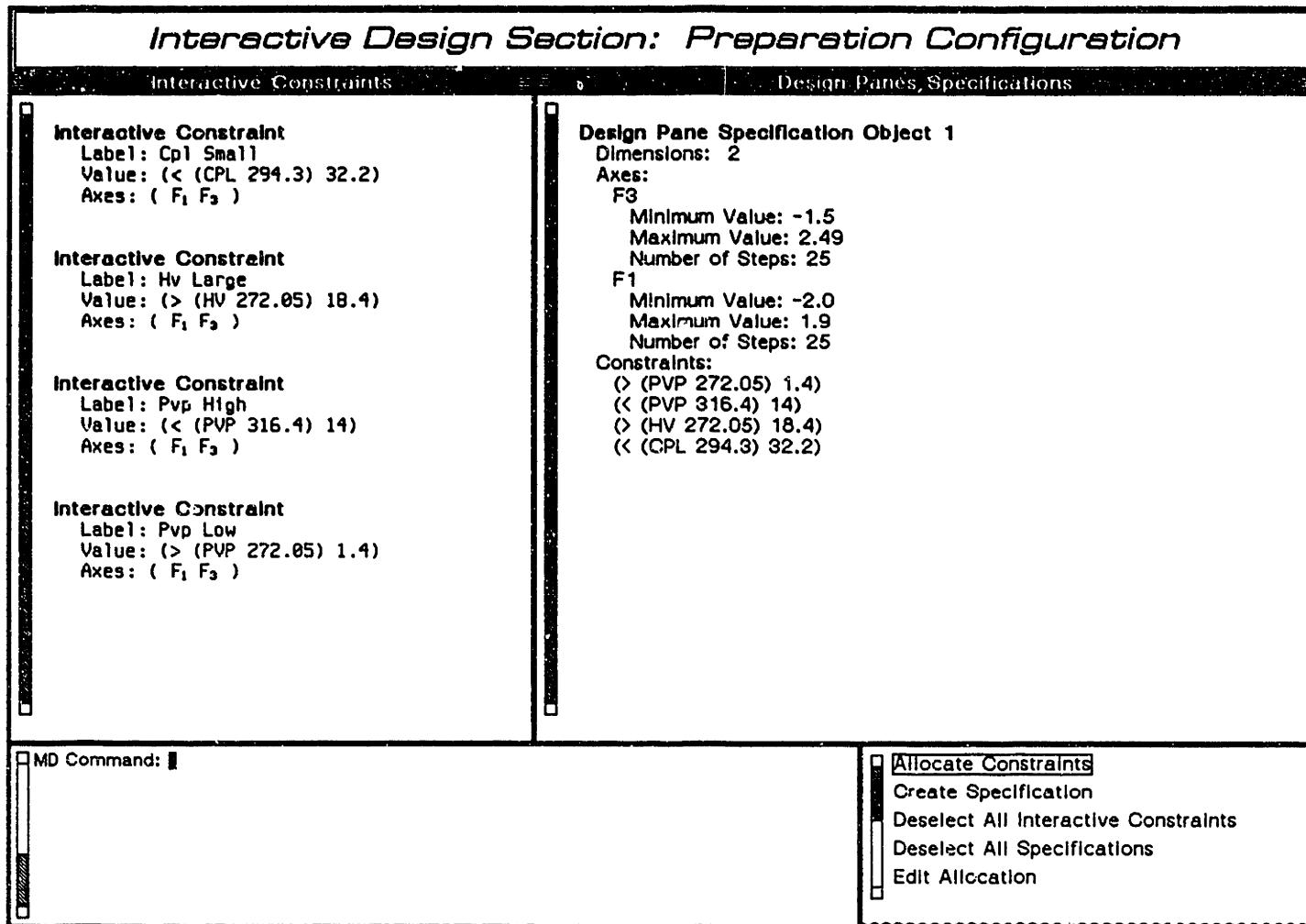


Figure 9.4: Interactive Constraint Allocation

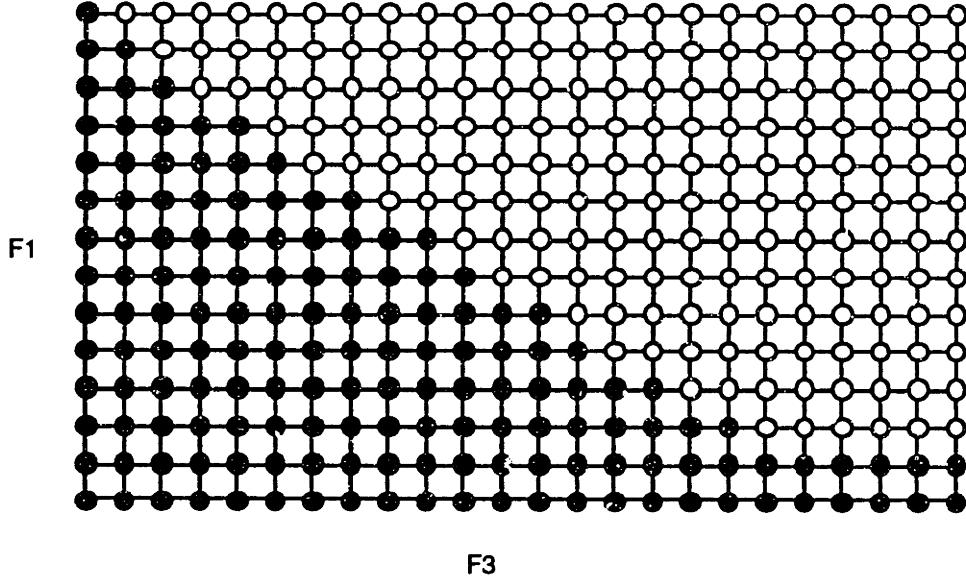


Figure 9.5: Interactive Design Space Solution Grid

range of values for each of the fundamental physical properties is divided into a set of points. The number of points is set by the designer with a default value of 25. Assuming the default value was used a *points-array* is constructed which is of dimensions  $25 \times 25$  with each cell containing a list of the form:

$$(x\text{-value} \quad y\text{-value}).$$

The constraint function for each of the interactive constraints allocated to a specification object is then evaluated at each point. The result of this evaluation is stored in an array equal in size to the discretization.

The Design Configuration presents the designer with one or more design spaces as a menu of groups to choose from. Figure 9.6 shows a design space formed from our example refrigerant constraints. Group vectors are sampled by moving the mouse over a group displayed in the Groups Menu Pane. Figure 9.7 shows a temporary  $-\text{CH}_3$

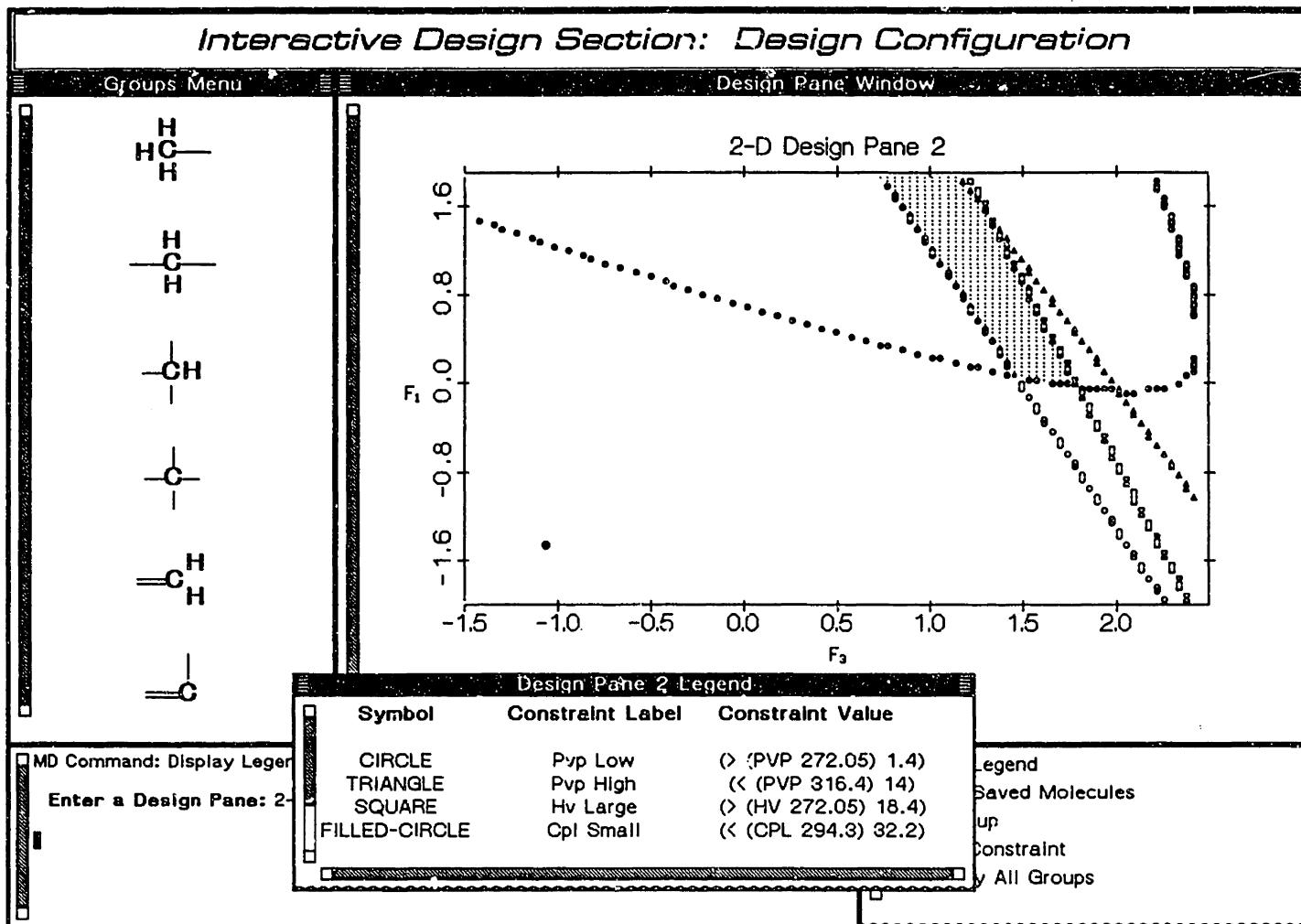


Figure 9.6: Legend for Design Space

group vector. Mouse **m-s-left** replaces the temporary group vector with a permanent one. Figure 9.8 shows a design space with the  $-\text{CH}_3$  group vector having been chosen and the  $-\text{Cl}$  group vector being sampled.

## 9.3 Preparation Configuration Objects

The two major objects used in the Preparation Configuration are:

1. **interactive-constraint**
2. **design-pane-specification-object**

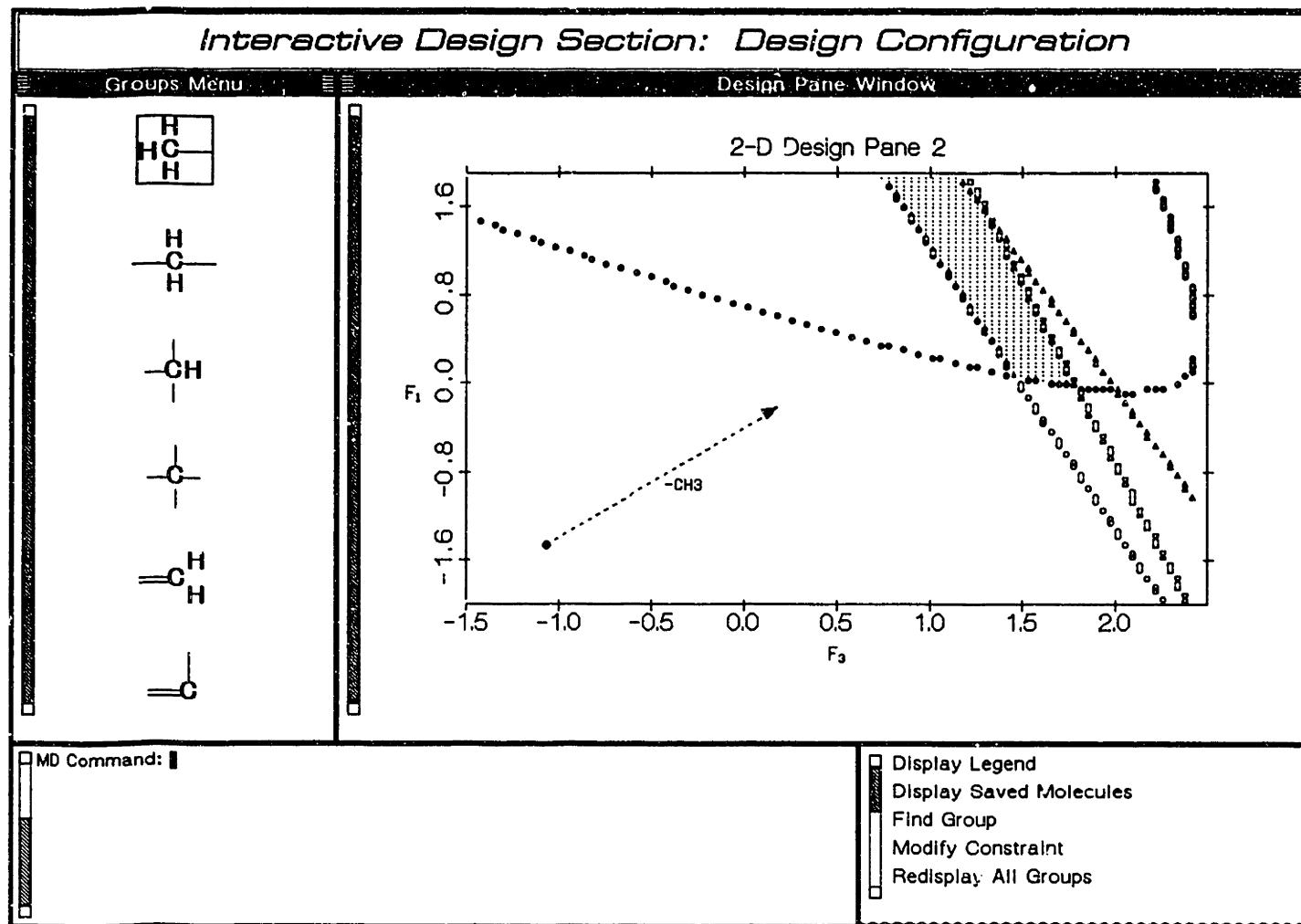
The definitions of these objects and their associated functions are in the file:

**molecular-design:interactive-design-section;objects.lisp.**

The Target Transformation Section's **Make Interactive** command transforms **transformed-constraints** into **interactive-constraints**. **Interactive-constraints** have four important instance variables:

1. **fundamental-properties**
2. **group-contribution-techniques**
3. **equation-oriented-techniques**
4. **constraint-function**

The group contribution and equation oriented estimation techniques stored in a **transformed-constraint** object are used to create a constraint function. This function estimates the needed physical properties to evaluate one physical property constraint. It returns **t** if the constraint is satisfied or **nil** if it is not. The code for the constraint function is stored in the **interactive-constraint**'s **constraint-function** instance

Figure 9.7:  $-CH_3$  Temporary Group Vector

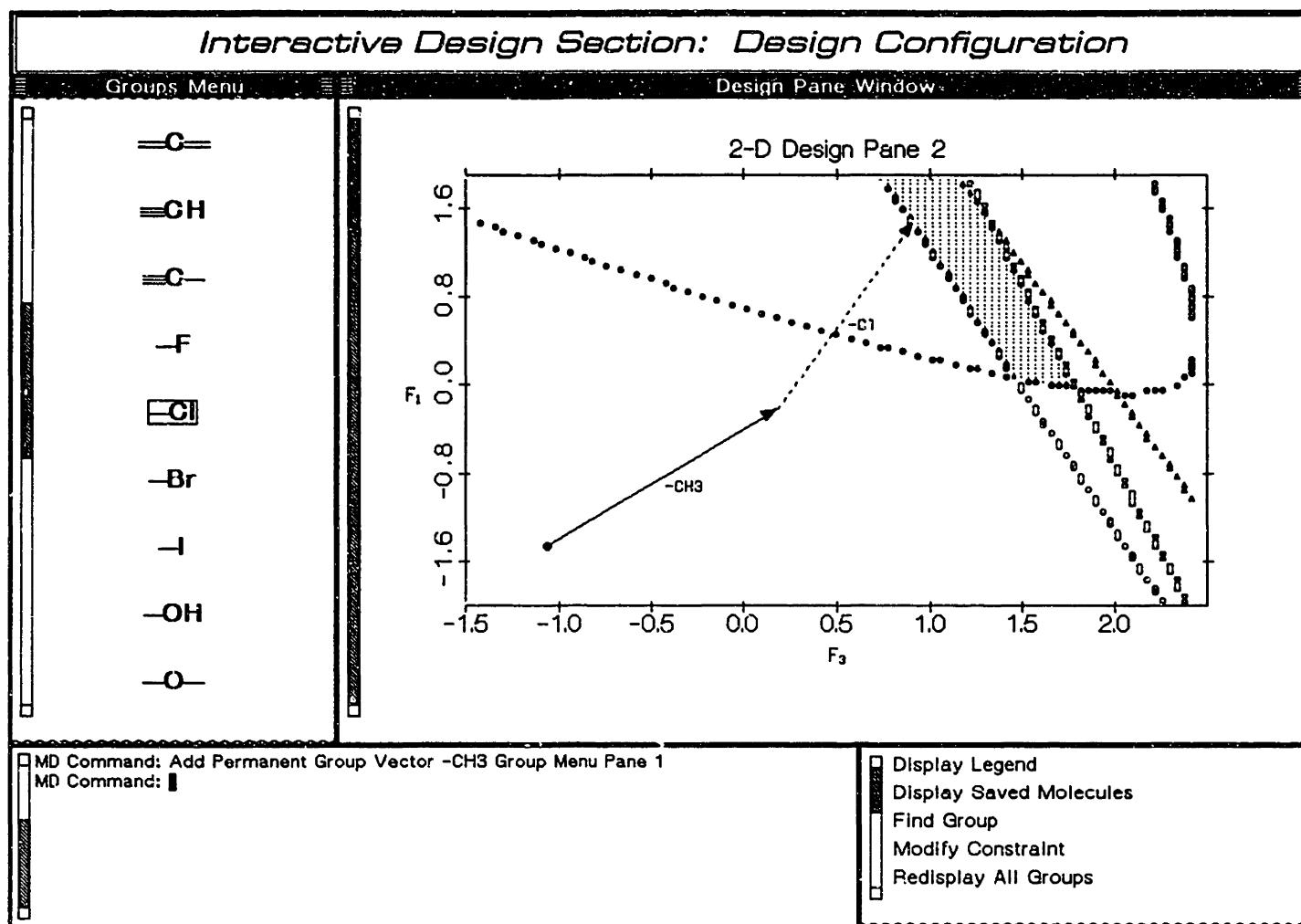


Figure 9.8: Molecule Within the Target Region

variable. The **fundamental-properties** instance variable stores the fundamental physical properties used as input to the constraint function. These fundamental properties become the axes of interactive design spaces.

The purpose of the **design-pane-specification-object** is to store the information needed to construct a design pane. The object has three major instance variables:

1. **property-specifications-list**
2. **number-of-divisions**
3. **interactive-constraints**

The **property-specification-list** stores a list of **property-specifications**. A **property-specification** is an object which stores a property, its minimum value, and its maximum value. This information is used in forming design panes. The **number-of-divisions** is used by the solution procedure to identify the feasible and infeasible regions for each of the constraints stored in the **interactive-constraints** instance variable.

## 9.4 Preparation Configuration Commands

The following commands are available in the Preparation Configuration's Command Menu. The command definitions are in the file:

```
molecular-design:interactive-design-section;commands.lisp.
```

**Allocate Constraints:** Adds the selected interactive constraints to each of the selected design pane specification objects.

**Create Specification:** Creates a specification object and adds it to the Design Panes Specifications Pane.

**Deselect all Interactive Constraints:** Deselects all the selected interactive constraints displayed in the Interactive Constraints Pane.

**Deselect all Specifications:** Deselects all selected specifications displayed in the Design Panes Specifications Pane.

**Edit Allocation:** This command prompts the designer to choose a design pane specification object. A menu is exposed querying the designer about keeping each constraint currently allocated to the chosen specification object. The command removes all constraints the designer chose not to keep. In this manner constraints mistakenly allocated can be easily removed.

**Form Design Panes:** Each selected design pane specification object creates a design space object. The physical property limits determine the extent of the design space's axes. Each of the allocated constraints are solved using the input step size and the procedure described in Section 9.2. The solved constraints are added to the design space object. Finally, the design space object is added to the Design Pane Window Pane of the Design Configuration and the system changes to the Design Configuration.

**Remove all Interactive Constraints:** Deletes all the interactive constraints displayed in the Interactive Constraints Pane.

**Remove all Specifications:** Deletes all design pane specification objects from the Design Panes Specifications Pane.

**Select all Interactive Constraints:** Selects all the unselected constraints displayed in the Interactive Constraints Pane.

**Select all Specifications:** Selects all the unselected design pane specification objects displayed in the Design Panes Specifications Pane.

**Specify Property Limits:** Prompts the designer for a design pane specification object. A menu is exposed querying the designer for the upper and lower bounds on physical property values and the number of steps to be used in the solution procedure. These values are queried for each fundamental physical property used in the allocated interactive constraints.

## 9.5 Design Configuration Commands

The following commands are available in the Design Configuration's Command Menu.

The command definitions are in the file:

`molecular-design:interactive-design-section;commands.lisp.`

**Display Legend:** Displays a table listing the symbol, constraint label, and constraint value for each of the design constraints displayed in a particular design pane. A design pane is prompted for and the table is displayed on a window resource exposed over the Design Pane Window Pane.

**Display Saved Molecules:** Displays all the saved design molecules known to the system on a window resource exposed over the Design Pane Window Pane. Design molecules are saved using the **Save Current Molecule** command.

**Find Group:** Prompts for the linear name of the group to be located. When a group is found the Groups Menu Pane scrolls so this group is located at the top of the window.

**Modify Constraint:** This command allows the designer to modify any of the constraints displayed in a design pane. Modification can involve manipulating the constraint value except changing the occurrences of the physical properties. This is because the physical property axes of the design pane correspond to the properties used in the constraint. Adding or deleting properties could necessitate a recalculation of the dimensionality and axes of the entire design space. I did not implement any features to accommodate so drastic a change.

Modifying a constraint by altering its values is a very powerful way to examine the sensitivity of the feasibility region to changes in property values. This is extremely important since design specifications are initially made without knowledge of their satisfiability.

This command first prompts for the design constraint which is to be modified. The designer should click on any one of the points which make up the constraint displayed in a design pane. The constraint's label and value are displayed allowing the designer to modify each. A typical modification of a constraint's value might be to change

$(> (P_{vp} 273) 1.01)$  (9.1)

to

$$(> (P_{vp} 273) 2.02). \quad (9.2)$$

Middle clicking on the displayed value allows the designer to edit the value. Left clicking will allow for the entry of a new value. It is recommended that the constraint's label also be changed to reflect the modification. This will keep the various graphical constraints organized and facilitate identification when displaying the legend of a design pane.

Once the constraint's have been modified a new constraint is formed and added to the design pane. I emphasize that the old constraint is not deleted. After the addition of a new constraint the feasibility region is recomputed and redisplayed.

**Redisplay all Groups:** A number of procedures change the groups which are displayed in the Groups Menu Pane. This command redisplays all groups which are consistent with the design panes displayed in the Design Pane Window Pane.

**Remove all Design Panes:** Deletes all the design pane objects from the Design Pane Window Pane.

**Remove Displayed Molecules:** All design molecules currently displayed in the design panes of the Design Pane Window Pane are removed. Displayed group vectors are not effected.

**Remove Group Vectors:** The displayed group vectors are removed from all the design panes displayed in the Design Pane Window Pane.

**Remove Saved Molecules:** The system maintains a list of all design molecules formed by the **Save Current Molecule** command. This command removes all these molecules.

**Save Current Molecule:** Forms a **design-molecule** from the current set of group vectors. This command is used once the appropriate set of group vectors are chosen to span the space from the starting point to the feasible region.

**Selection Assist:** This command provides a menu containing commands to assist in the selection of groups. The commands fall into two categories: *Groups Display* and *Sorting*.

The **Groups Display** category has two options:

1. **Display all Groups:** Displays the group vectors for all the groups displayed in the Groups Menu Pane in a chosen design pane.
2. **Choose Angle:** Limits the groups displayed in the Groups Menu Pane to those whose group vectors satisfy specified angle constraints. The command first prompts for a design pane. The command then prompts the designer to specify the angle with the mouse. Those groups whose group vectors do not fall within the specified angle are removed from the Groups Menu Pane.

The **Sorting** category has four options:

1. **Sort by Magnitude:** The command first prompts for a design pane. The groups displayed in the Groups Menu Pane are then sorted with respect to the magnitude of their group vectors displayed in the chosen design pane. The Groups Menu Pane is redisplayed with the groups ordered from smallest to largest magnitude.
2. **Sort by Angle:** The command first prompts for a design pane. The groups displayed in the Groups Menu Pane are then sorted with respect to the angle of their group vectors displayed in the chosen design pane. The Groups Menu Pane is redisplayed with the groups ordered from smallest to largest angle.
3. **Sort by Closest Distance:** The command prompts for a group vector. The groups displayed in the Groups Menu Pane are sorted with respect to the distance between their group vectors and the one chosen. The Groups Menu Pane is redisplayed with the groups ordered from closest to furthest.

4. **Sort by Closest Angle:** The command prompts for a group vector. The groups displayed in the Groups Menu Pane are sorted with respect to the angle between their group vectors and the one chosen. The Groups Menu Pane is redisplayed with the groups ordered from closest to furthest.

**Specify Groups:** The groups displayed in the Groups Menu Pane can be restricted using a variety of criteria. This command exposes a menu displaying the following criteria:

**Atoms Screening**

    Disallowed Atoms    Required Atoms

**Ring Class**

    Cyclic    Acyclic    Mixed

**Global Valence**

    One    Two    Three    Four

**Free Bond Types**

Single	Double	Triple
Ring-Single	Ring-Double	Ring-Triple

Groups not satisfying any of the criteria are removed from the Groups Menu Pane.

## 9.6 Section Discussion

The Preparation Section is a necessary intermediate step. However, many of the facilities provided by the Preparation Configuration could be condensed into the Design Configuration. This requires improving the graphic functionality of design panes. The ability to add new constraints to existing design panes, rescale, and zoom would facilitate the design process. Three dimensional design spaces are also necessary.

Interfacing the Design Configuration with the physical property data base would enable existing compounds to be displayed in a design pane. The existing compounds

could be “brought into” the design pane. This would allow existing compounds to be used as the basis of evolutionary designs.

The solution procedure for interactive constraints can be significantly improved. To present the feasible region and constraint edges it is necessary that all constraints be solved at the same points. The existing solution procedure accomplished this. Running the solution procedure in a hierarchical manner would improve speed. The first discretization would identify rectangles in which an edge occurs. This rectangle would then be further discretized and the solution procedure repeated.

## 9.7 Example Usage

We design refrigerants in this example. Two tasks must be performed: 1) entering design pane specifications; 2) choosing appropriate group vectors. The Preparation Configuration accomplishes the task of entering design pane specifications. The Design Configuration displays the design panes, groups, and group vectors enabling the designer to create molecules.

### Entering Property Constraints

We enter four constraints on physical properties important to refrigerants. These constraints are entered in the Problem Formulation Section.

**Action 9.1** *Mouse right on a empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 9.2** *Mouse left on the Problem Formulation Section Constraints Configuration.*

The system changes configuration to the Problem Formulation Section. Enter the following four constraints:

(> (Pvp 272.05) 1.4)

(< (Pvp 316.4) 14)

(> (Hv 272.05) 18.4)

(< (Cpl 294.2) 32.2)

The details of this entry is discussed in the Problem Formulation Section Example Usage, Section 4.6.

### **Transforming Constraints**

Our goal in transforming constraints for interactive design is to develop estimation procedures which require two fundamental physical properties. Only the selected constraints are transformed.

**Action 9.3** *Mouse left on the Select All Constraints command.*

All the constraints displayed in the Constraints Pane are selected.

**Action 9.4** *Mouse left on the Verify Constraints command.*

The system reports that all constraints were verified.

**All constraints were verified.**

**Action 9.5** *Mouse left on the Transform Constraints command.*

The system coerces each of the constraints into a transformed constraint. These transformed constraints are added to the Target Transformation Section's Transformed Constraints Pane. The system then changes to the Target Transformation Section.

**Action 9.6** *Mouse left on the Select All Transformed Constraints command.*

All the transformed constraints displayed in the Transformed Constraints Pane are selected.

**Action 9.7** *Mouse h-sh-left on the Acentric-Factor ( $P_c T_{br}$ ) – Lee Kesler Acentric Factor -- Equation Oriented Technique.*

**Action 9.8** *Mouse h-sh-left on the  $P_c \rightarrow (F_1 F_2 F_3)$  – Joback  $P_c$  Factor – Equation Oriented Technique.*

**Action 9.9** *Mouse h-sh-left on the  $H_{vb} \rightarrow (F_1 F_2 F_3)$  – Joback  $H_{vb}$  Factor – Equation Oriented Technique.*

**Action 9.10** *Mouse h-sh-left on the  $H_v \rightarrow (\Delta H_{vb} T_c T_b)$  – Watson Relation  $T_c$  Biased – Equation Oriented Technique.*

**Action 9.11** *Mouse h-sh-left on the  $F2 \rightarrow ()$  –  $F2$  Assumption – Equation Oriented Technique.*

**Action 9.12** *Mouse h-sh-left on the  $F3 \rightarrow ()$  – Joback  $F3$  – Group Contribution Technique.*

**Action 9.13** *Mouse h-sh-left on the  $F_1 \rightarrow ()$  – Joback  $F_1$  – Group Contribution Technique.*

**Action 9.14** *Mouse h-sh-left on the  $Cpl \rightarrow (Cpv \text{ omega } Tc)$  – Rowlinson  $Cpl$  – Equation Oriented Technique.*

**Action 9.15** *Mouse h-sh-left on the  $Cpv \rightarrow (F_1 \ F_2 \ F_3)$  – Joback  $Cpv$  298 Factor – Equation Oriented Technique.*

**Action 9.16** *Mouse h-sh-left on the  $Tc \rightarrow (F_1 \ F_2 \ F_3)$  – Joback  $Tc$  Factor – Equation Oriented Technique.*

**Action 9.17** *Mouse h-sh-left on the  $Tbr \rightarrow (Tb \ Tc)$  –  $Tbr$  Definition – Equation Oriented Technique.*

**Action 9.18** *Mouse h-sh-left on the  $Tb \rightarrow (F_1 \ F_2 \ F_3)$  – Joback  $Tb$  Factor – Equation Oriented Technique.*

**Action 9.19** *Mouse h-sh-left on the  $Pvp \rightarrow (Tb \ Tc \ Pc)$  – Riedel Plank Miller  $Tc$  Biased – Equation Oriented Technique.*

**Action 9.20** *Mouse left on the **Apply Selected Techniques** command.*

All transformed constraints contain only two fundamental physical properties and no non-fundamental physical properties.

**Action 9.21** *Mouse left on the **Check Fundamentality** command.*

The system reports that all constraints were verified:

**All Selected Constraints have been Reduced to Fundamental Properties.**

**Action 9.22** *Mouse left on the Check Dimensionality command.*

The system reports that all constraints were verified.

**All Selected Constraints are of Dimension 3 or Less.**

## Design Pane Preparation

Constraints are entered into the Preparation Configuration only from the Target Transformation Section.

**Action 9.23** *Mouse left on the Make Interactive command.*

The system checks each of the transformed constraints for dependency only on fundamental physical properties and for a dimensionality of 3 or less. If all the transformed constraints satisfy these checks the system prints messages noting this:

**All Selected Constraints are of Dimension 3 or Less.**

**All Selected Constraints have been Reduced to Fundamental Properties.**

The system then coerces each of the transformed constraints into interactive constraints, adds these constraints to the Preparation Configuration's Interactive Constraints Pane, and changes to the Preparation Configuration.

The specifications of a design pane are entered into a `design-pane-specification` object. Specifications include:

1. The constraints to be displayed in the design space.
2. The axes of the design space.
3. The upper and lower bounds on each of the axes.
4. The number of divisions used in solving the constraints.

We begin by creating a **design-pane-specification** object.

**Action 9.24** *Mouse left on the Create Specification command.*

A **design-pane-specification** object is created and added to the Specifications Pane.

The new specifications object contains no constraints and thus has no information on axes or axes limits.

We now specify which constraints are to be displayed in the design space to be created. In this example we add all four of our refrigerant design constraints to the specifications object.

**Action 9.25** *Mouse left on the Select All Interactive Constraints command.*

**Action 9.26** *Mouse left on the Select All Specifications command.*

**Action 9.27** *Mouse left on the Allocate Constraints command.*

The **Allocate Constraints** command collects all of the selected constraints in the Interactive Constraints Pane and adds them to each of the selected specifications object displayed in the Specifications Pane.

The system determines the axes of the design pane to be created. Each physical property known to the system has a default minimum and maximum value specified. These are displayed in the specifications object. The number of steps used to solve the constraints defaults to 25.

To improve the appearance of the design space we increase the number of steps from 25 to 80.

**Action 9.28** *Mouse left on the Specify Property Limits command.*

The system prompts for a specification object:

**Enter a Specification Object:**

**Action 9.29** *Mouse left on the Design Pane 1 design pane specifications object.*

The system exposes a menu prompting for the minimum, maximum, and number of steps for each of the physical property axes.

**Action 9.30** *Mouse left on the number 25 appearing after the Number of Steps: prompt under the F1 axis.*

The number changes to a blinking cursor awaiting input.

**Action 9.31** *Type in the number 80. Press the return key when the entry is complete.*

The process is repeated for the F3 axis.

**Action 9.32** *Mouse left on the number 25 appearing after the Number of Steps: prompt under the F3 axis.*

The number changes to a blinking cursor awaiting input.

**Action 9.33** *Type in the number 80. Press the return key when the entry is complete.*

**Action 9.34** *Press the ENTER key when both entries are complete.*

The number of steps for both axes is replaced in the design pane specifications object.

The Specifications Pane is redisplayed to show the new values.

Now that the specifications have been entered the system constructs a design pane.

## Designing Interactively

Once the specifications are entered into a design pane specifications object the next step is to use these specifications to create a design space. The system creates a design space for each of the design pane specifications object selected in the Specifications Pane.

**Action 9.35** *Mouse left on the Select all Specifications command.*

**Action 9.36** *Mouse left on the Form Design Panes command.*

The system creates a design pane, adds this pane to the Design Pane of the Design Configuration, and changes to the Design Configuration. Specifying an  $80 \times 80$  grid causes the system to take approximately 5 minutes to solve the constraints.

Once the system has solved the constraints it forms a design pane, changes to the Design Configuration, and adds the design pane to the Design Pane Window Pane. The groups for each of the fundamental physical properties composing the design panes axes are intersected. The resulting intersection of groups is displayed in the Groups Menu Pane.

We resize the design pane.

**Action 9.37** *Mouse left on the Resize Design Pane command.*

The system prompts for a design pane to be resized:

**Enter a design pane to be resized:**

**Action 9.38** *Mouse left on the 2D-active-graph displayed in the Design Pane Window Pane.*

The mouse cursor moves to the Design Pane Window Pane and changes into an upper-left corner.

**Action 9.39** *Mouse left near the upper left corner of the Design Pane Window Pane.*

This positions the new upper left corner of the design space. The mouse cursor now appears as a lower right corner. As you move the mouse a “rubber-banding box” is drawn connecting the affixed upper left corner and the lower right corner mouse cursor.

This box roughly shows the new size of the design space.

**Action 9.40** *Mouse left near the lower right corner of the Design Pane Window Pane.*

The constraints are denoted on the design pane using geometric symbols. A legend explains the meaning of each symbol.

**Action 9.41** *Mouse left on the Display Legend command.*

The system prompts for a design pane:

**Enter a Design Pane:**

**Action 9.42** *Mouse left on our design pane displayed in the Design Pane Window Pane.*

The system displays a legend detailing the symbol, constraint name, and constraint value for each of the constraints displayed in our design pane.

**Action 9.43** *Press the space bar.*

Group vectors are “sampled” by placing the mouse over a group displayed in the Groups Menu Pane.

**Action 9.44** *Move the mouse over the  $-CH_3$  group displayed in the Groups Menu Pane.*

A temporary group vector is drawn on our design pane. Temporary group vectors are denoted by dashed arrows. Figure 9.7 showed what the screen displays should resemble. If we had multiple design panes, the group vector would be drawn on all design panes.

To include a group in a design molecule the designer chooses the group.

**Action 9.45** *Mouse m-s-left on the  $-CH_3$  group displayed in the Groups Menu Pane.*

The temporary group vector is replaced by a permanent group vector. Permanent group vectors are denoted by a solid arrow. The origin of group vectors now moves to the head of our  $-CH_3$  group vector.

**Action 9.46** *Mouse m-s-left on the  $-Cl$  group displayed in the Groups Menu Pane.*

Choosing this second group forms a design molecule which satisfies our constraints.

**Action 9.47** *Mouse left on the Save Current Molecule command.*

This resets the origin allowing a new design to begin. Displaying several molecules simultaneously is confusing.

**Action 9.48** *Mouse left on the Remove Displayed Molecules command.*

# Chapter 10

## Molecule Evaluation Section

The Evaluation Section provides facilities for estimating molecules' physical properties. Estimating physical properties is especially useful when formulating the design target and evaluating designed molecules. During problem formulation we may need to estimate the properties of compounds currently in use. In final evaluation we would like to have the property profile of the designed compounds available for inspection.

One of the major objectives of the Evaluation Section is to provide estimation techniques of the highest accuracy. These estimation techniques may not be appropriate for use in the design procedures. They would thus serve as an additional check to verify the efficacy of any molecules designed.

The Evaluation Section is divided into two configurations: 1) Specifications Configuration; 2) Values Configuration. The main task of the Specifications Configuration is to provide facilities for the creation of an estimation procedure. The Values Configuration applies these estimation procedures to input molecules.

## 10.1 Section Layout

The screen layout of the Specifications Configuration is shown in Figure 10.1. The screen real estate is used by six panes:

**Evaluation Section Specifications Configuration Title Pane:** Displays the title of the Specifications Configuration.

**Procedure Development Pane:** This pane is the major focus of the Specifications Configuration. Estimation procedures are displayed in this pane in a decision tree representation. The designer creates an estimation procedure by choosing from the displayed estimation techniques. Estimation procedures are developed in a tree-like manner.

**Documentation Entry Pane:** This is a Zwei editor used to enter documentation for newly created estimation procedures. This pane should be replaced in future versions of the software by an md-editor resource.

**Procedures Pane:** This pane displayed all the estimation procedures known to the system. When an estimation procedure is created it is given a “pretty-name”. It is this pretty name which is displayed.

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

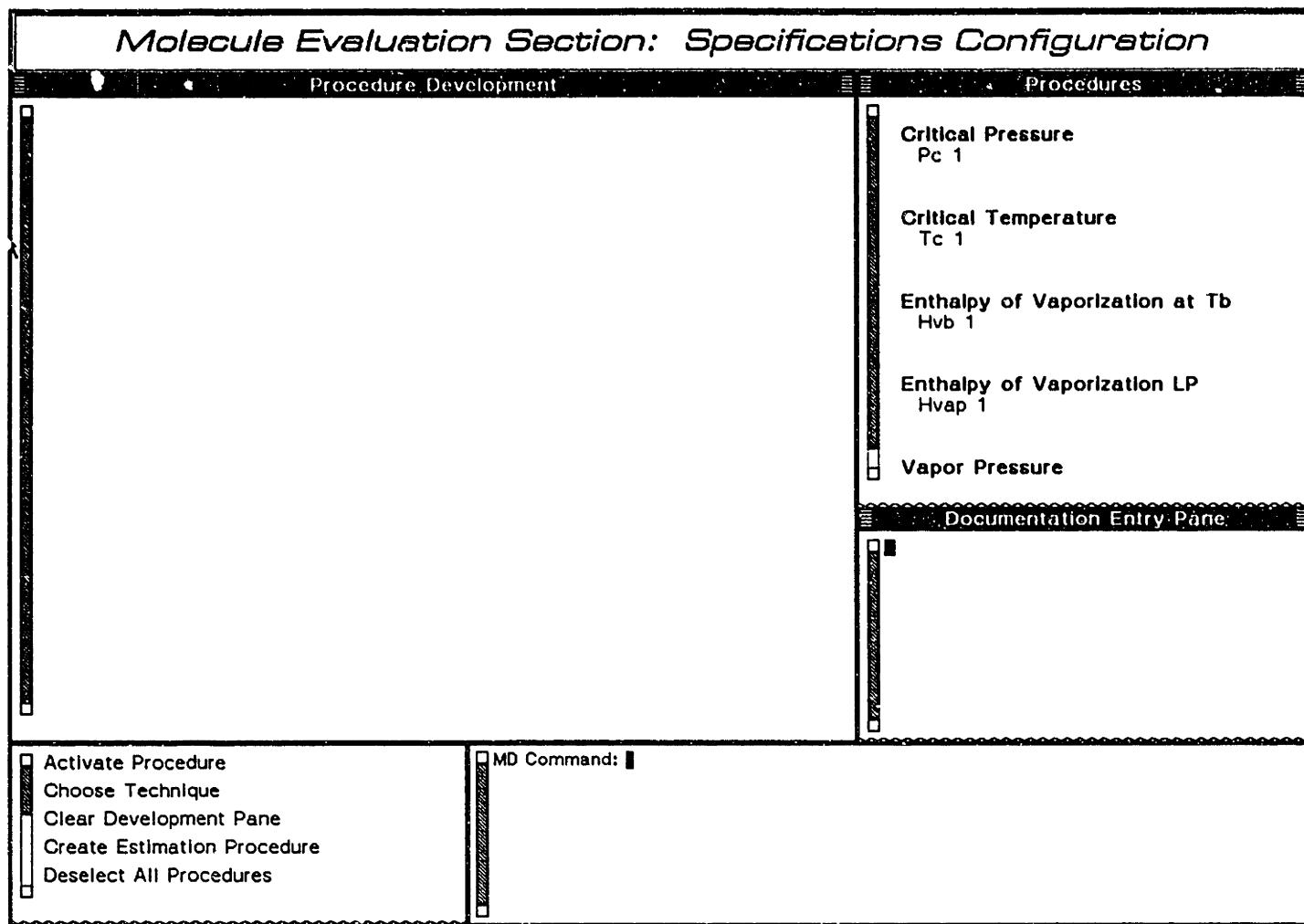


Figure 10.1: Evaluation Section Specifications Configuration Screen

**Evaluation Section Specifications Configuration Commands Menu:** The command menu containing commands relevant to the Specifications Configuration.

The screen layout of the Values Configuration is shown in Figure 10.2. The screen real estate is used by six panes:

**Evaluation Section Values Configuration Title Pane:** Displays the title of the Values Configuration.

**Selection Choice Pane:** This pane is an AVV-pane. Its purpose is to allow the designer to specify how physical properties are estimated. The idea was that for a particular physical property,  $P_{vp}$  for example, there might be several estimation procedures. One of these would be designated as the default estimation procedure. If a molecule's  $P_{vp}$  was being estimated and **Default** was selected in the Selection Choice Pane then the default  $P_{vp}$  estimation procedure would be used. If **Prompt** was selected in the Selection Choice Pane the designer would be requested to choose from all the estimation procedures available for  $P_{vp}$ . The facility is not currently operational. The system is in **Prompt** mode.

**Estimated Values Pane:** This pane is the major focus of the Value Configuration. Molecules are displayed in this pane along with any of their estimated physical properties.

**Groups Pane:** Molecules can be entered in by selecting a collection of groups from those displayed in the Groups Pane.

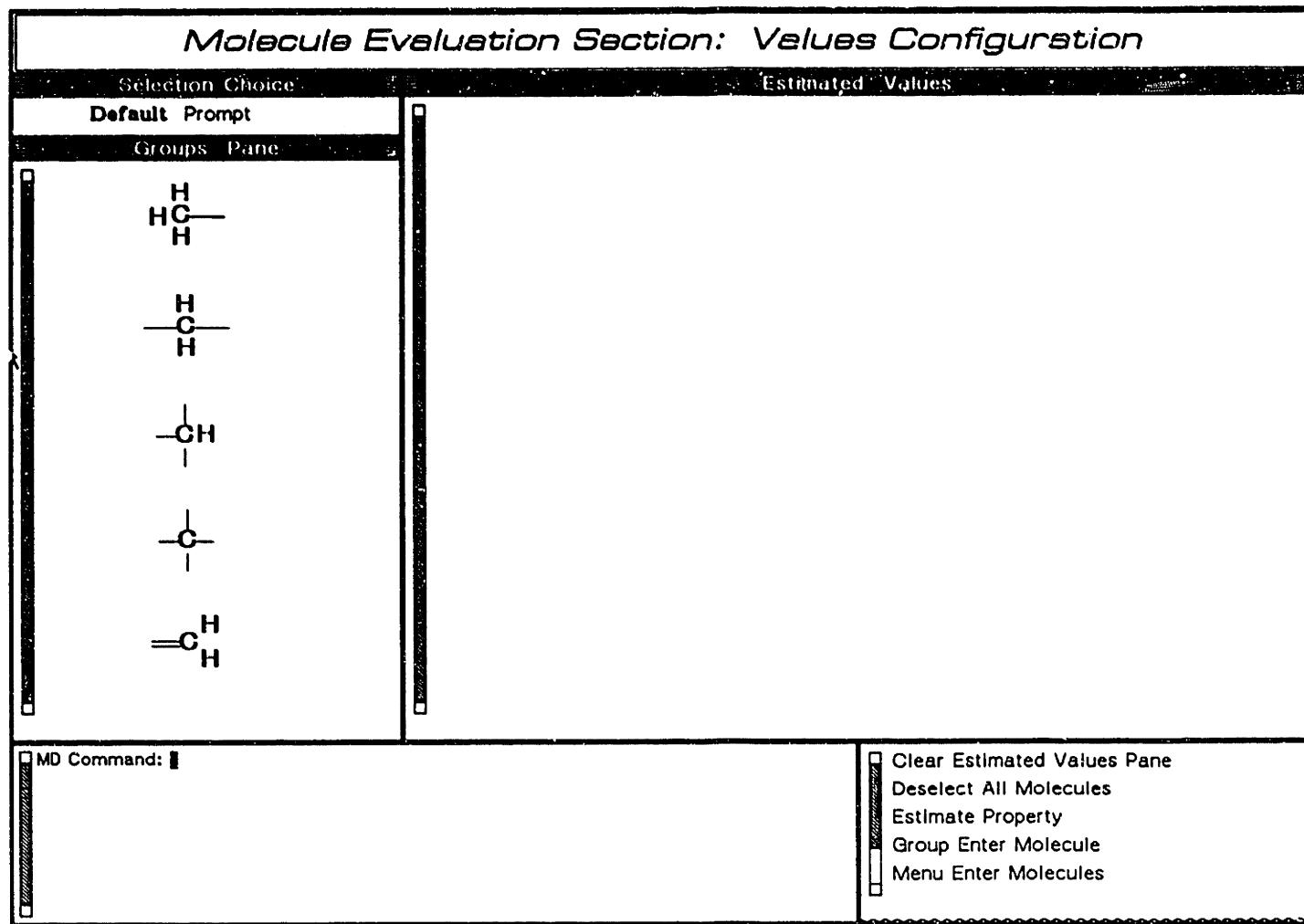


Figure 10.2: Evaluation Section Specifications Configuration Screen

**Molecular Design Interaction Pane:** The interactor pane for accepting commands and input.

**Evaluation Section Values Configuration Commands Menu:** The command menu containing commands relevant to the Values Configuration.

## 10.2 Section Operation

The Evaluation Section addresses two tasks: 1) development of estimation procedures from estimation techniques known to the system; 2) estimation of molecules' physical properties using these estimation procedures. The development of estimation procedures is done by the Specifications Configuration. The estimation of physical properties is done by the Values Configuration.

### 10.2.1 Estimation Procedure Development

Estimation procedures estimate a molecule's physical properties given its molecular structure and temperature and pressure if needed. Developing a new estimation procedure occurs in the Specifications Configuration. The **Create Estimation Procedure** command prompts for the physical property to be estimated and the name of the new procedure. Once this information has been entered the system creates an estimation procedure graph object which is displayed in the Procedures Development Pane. Figure 10.3 shows the initial display of a new estimation procedure.

Using the **Choose Technique** command we select the estimation techniques which

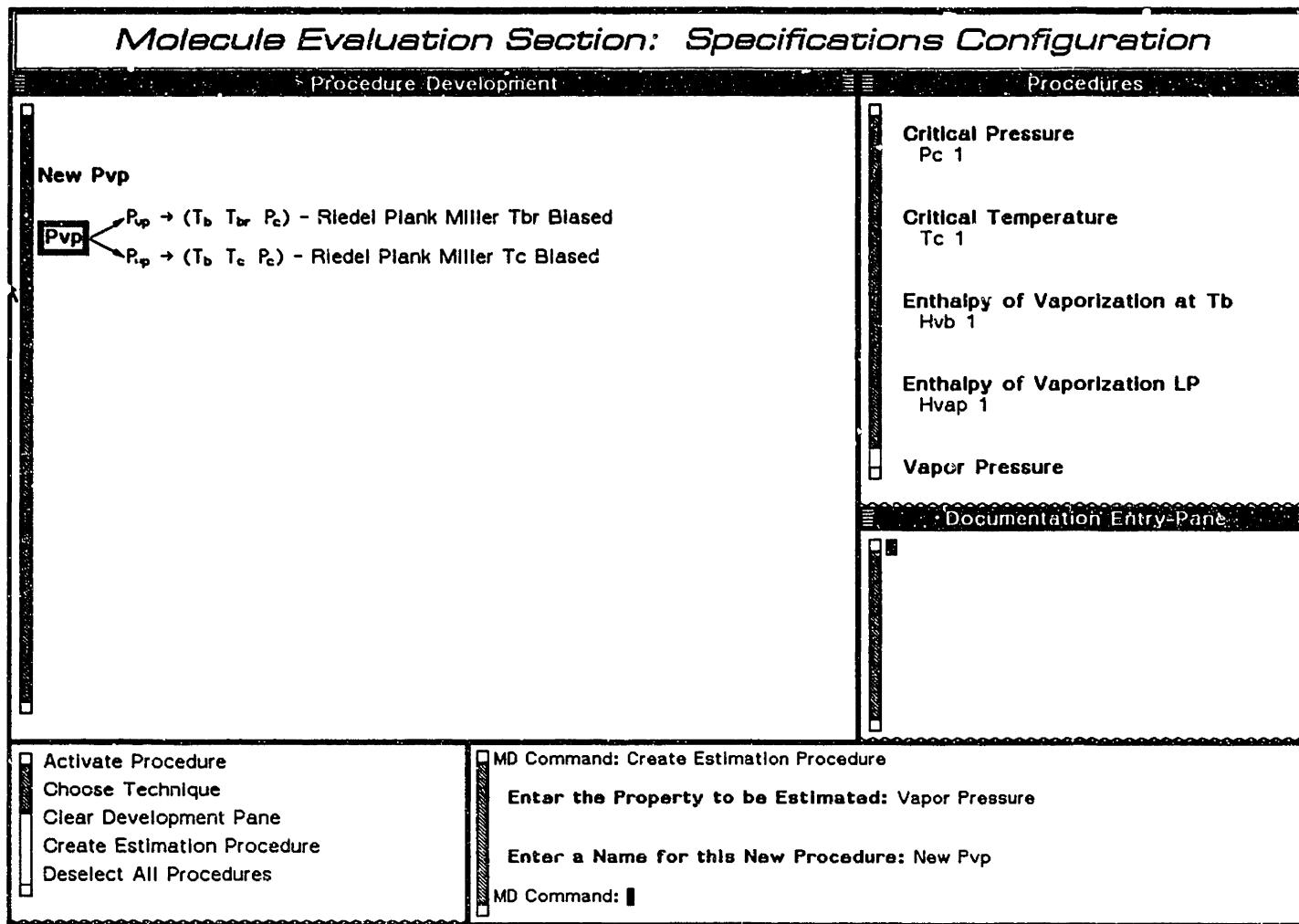


Figure 10.3: Display of Initial  $P_{vp}$  Estimation Procedure

compose the estimation procedure. The choice of one estimation technique may necessitate the choice of another. Figure 10.4 shows the state of our estimation procedure after choosing the Riedel-Plank-Miller Tbr EOT to estimate  $P_{vp}$ . Choosing estimation techniques continues until all of the leaf physical properties are specified. Figure 10.5 shows the final state of our new estimation procedure. The **Document Technique** command provides information useful in deciding between estimation techniques. The **Respecify Property** command enables the designer to change his or her estimation technique choice.

The **Activate Procedure** command translates the decision tree specified by our estimation procedure graph into a LISP function. This function is stored in an estimation procedure object and displayed in the Procedures Pane. The **Edit Procedure Documentation** command allows the designer to enter documentation for the newly created procedure.

### 10.2.2 Physical Property Estimation

The estimation procedures developed in the Specifications Configuration are used to estimate the physical properties of molecules. Molecules displayed in the Estimated Values Pane are capable of having their properties estimated. A molecule can be entered into the Estimated Values Pane in two ways. The **Menu Enter Molecule** command displays all the estimated molecules known to the system in a window resource exposed over the Estimated Values Pane. Estimated molecules are created by the **Save Current Molecule** of the Interactive Design Section's Design Configuration. Choosing a molecule from this display enters it into the Estimated Values Pane. The

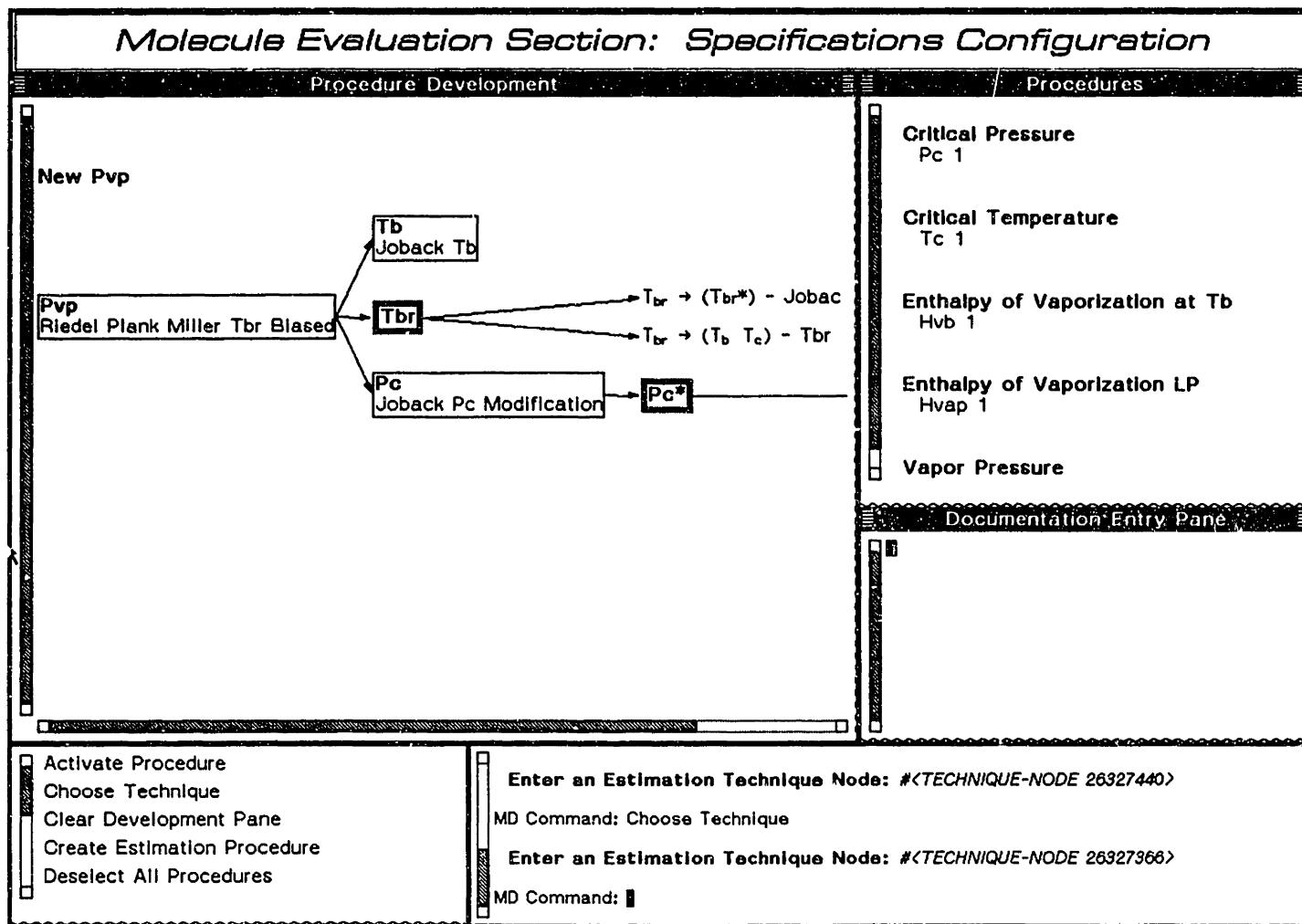


Figure 10.4: Display of Intermediate  $P_{vp}$  Estimation Procedure

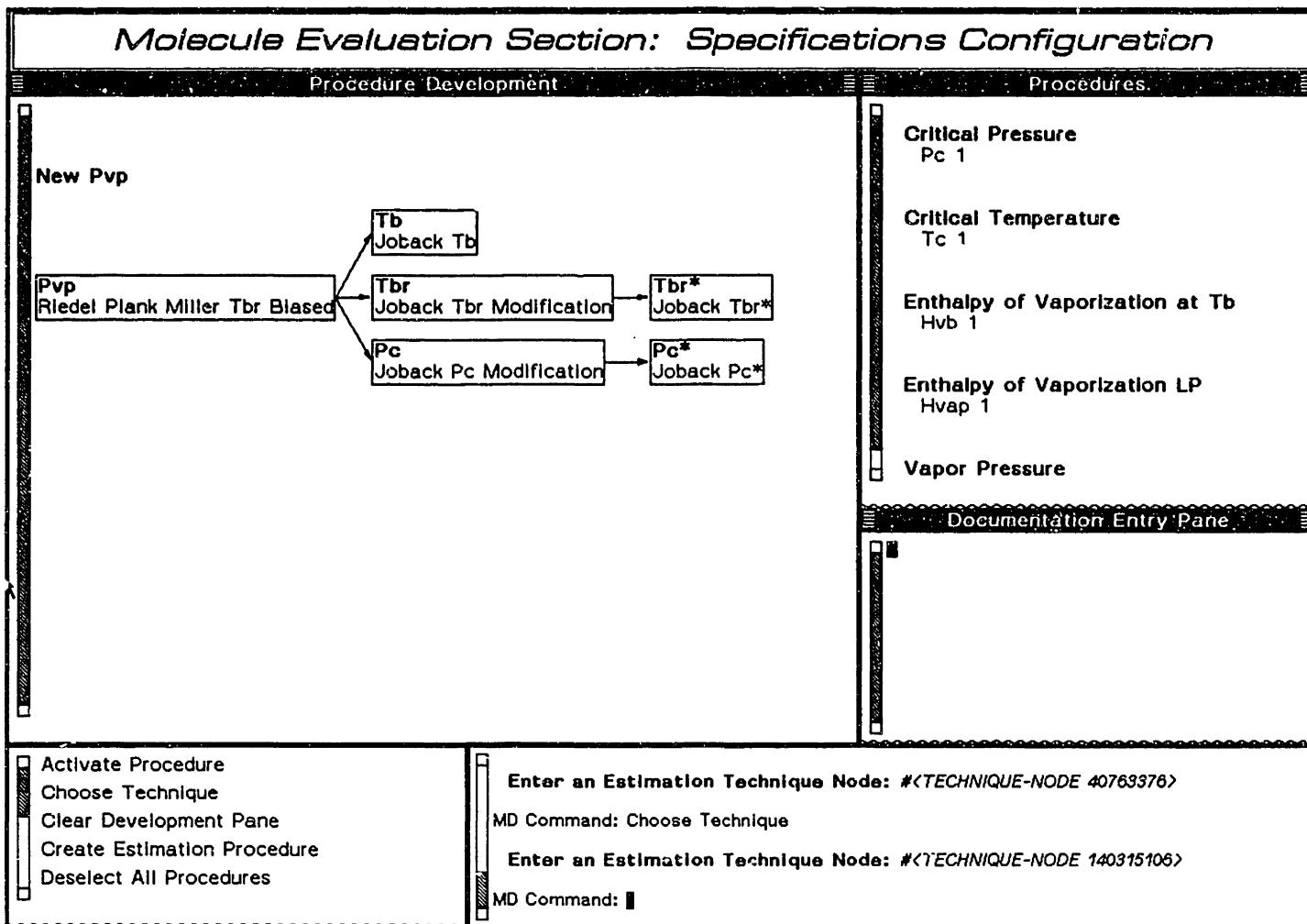


Figure 10.5: Display of Final  $P_{vp}$  Estimation Procedure

**Group Enter Molecule** command prompts the designer for a sequence of groups. The groups are collected and formed into an estimated molecule object. This molecule is then added to the Estimated Values Pane.

Once a molecule has been entered into the Estimated Values Pane its physical properties can be estimated. The **Estimate Property** command prompts the designer for a property. If this property is dependent upon state variables, e.g.  $P_{vp}$  is dependent upon temperature, then a sequence of their values is prompted for. The **Estimate Property** command estimates properties only for the selected molecules displayed in the Estimated Values Pane. Selection of molecules is done by done using a **shift-hyper-left** mouse gesture. Each molecule displayed its estimated properties. Figure 10.6 shows  $P_{vp}$  estimated at the temperatures: 250, 275, 300, 325, 350, 375, and 400, for dichlorodifluoromethane.

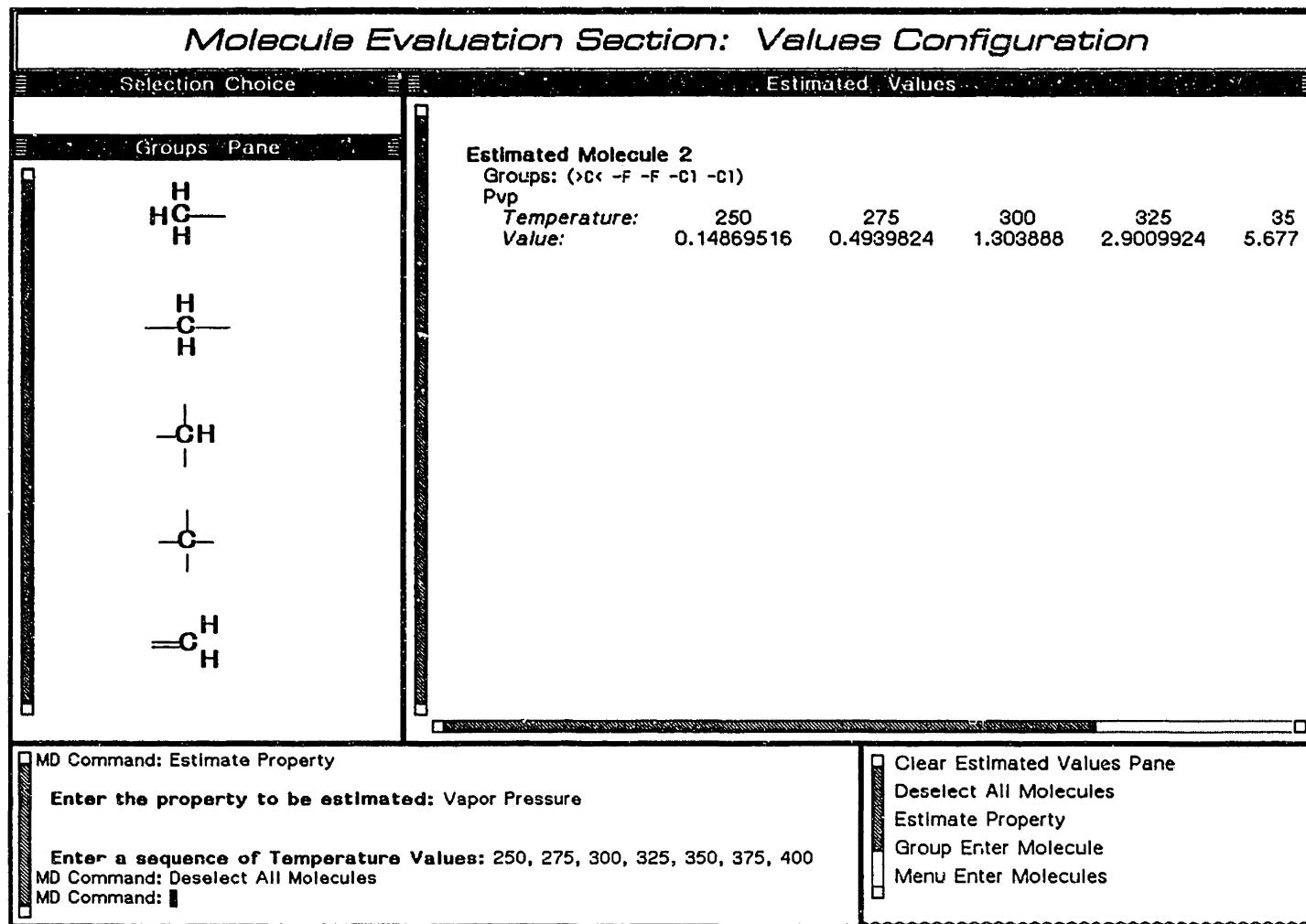
### 10.3 Specifications Configuration Objects

The six major objects used in the Specifications Configuration are:

1. **graph-node**
2. **technique-node**
3. **value-node**
4. **property-node**
5. **estimation-procedure-graph**
6. **estimation-procedure**

The definitions for these objects and their associated functions are in the file:

`molecular-design:evaluation;objects.lisp.`

Figure 10.6: Estimated  $P_{vp}$  for Dichlorodifluoromethane

The instance variables and important functionality for each of these objects is discussed.

The **graph-node** object is a base object on which the **technique-node**, **value-node**, **property-node**, and **estimation-procedure-graph** objects are built. It provides two instance variables:

1. **parent**
2. **children**

which store the objects composing the estimation procedure graph. The **display-graph** method uses the information stored in these instance variables to display the graph.

**Value-node** and **technique-node** objects do not have children. The **value-node** contains the single instance variable: **value**, which contains the physical property's value specified by the designer. The **technique-node** contains the single instance variable: **technique**, which contains the estimation technique object chosen by the designer. Each node object has its own **md-present-self** method.

**Property-node** objects contain two instance variables:

1. **property**
2. **technique**

The physical property being estimated is stored in the **property** instance variable. The estimation technique chosen by the designer is stored in the **technique** instance variable.

The three major tasks a **property-node** object must perform are: 1) presenting itself; 2) storing the chosen estimation technique or value; 3) creating LISP code to perform the estimation. The **Choose Technique** calls the **update-property-node** method. This method stores the designer's choice for an estimation technique and updates the appropriate nodes. The **activate-self** method creates the LISP function

which estimates the property by the chosen techniques. The `md-present-self` method presents the object differently depending upon whether or not an estimation technique has been selected.

The `estimation-procedure-graph` object stores the graph's starting node. It is the `estimation-procedure-graph` object which is displayed in the Procedure Development Pane. The `estimation-procedure` object has two instance variables:

1. `estimated-property`
2. `estimation-function`

which store the information needed for physical property estimation. The `save-to-file` method of the `estimation-procedure-graph` object writes the estimation procedure to file.

## 10.4 Values Configuration Objects

The two major objects used in the Values Configuration are:

1. `estimated-molecule`
2. `estimated-property`

The definitions for these objects and their associated functions are in the file:

`molecular-design:evaluation;objects.lisp`.

The instance variables and important functionality for each of these objects is discussed.

The `estimated-molecule` object has two instance variables:

1. `groups`
2. `property-values`

The `groups` instance variable stores the groups which make up the molecule. The `property-values` instance variables contains a list of `estimated-property` objects.

`Estimated-property` objects estimate physical property values. Each `estimated-property` object estimates a specific property for a specific molecule. The object is presented in a variety of ways depending upon the presence or absence of state variable dependence.

## 10.5 Specifications Configuration Commands

The following commands are available in the Specifications Configuration's Command Menu. The command definitions are in the file:

```
molecular-design:evaluation;commands.lisp.
```

**Activate Procedure:** Once estimation techniques are chosen for all required physical properties this command begins the process of forming the LISP code which implements the estimation procedure. The command first prompts for an estimation procedure graph. This graph is chosen from those displayed in the Procedure Development Pane. The command verifies that all nodes were specified. The LISP function is then formed. This function is stored in an estimation procedure object along with the estimated physical property and the procedure's name. Finally the new estimation procedure object is added to the Procedures Pane.

**Choose Technique:** During procedure development the system collects applicable estimation techniques for each property of the estimation procedure. These estimation

techniques are displayed as the children of the property being connected to the property node by arrows. To specify a particular estimation technique the designer first selects this command and then chooses the desired technique. The estimation procedure display is updated to show the property is estimated by the chosen technique.

**Clear Development Pane:** This removes all the objects from the Procedure Development Pane.

**Create Estimation Procedure:** This command begins the process of developing a new estimation procedure. The designer is first prompted for the physical property the new estimation procedure will estimate. All the properties known to the system are displayed in a window resource exposed over the Procedures Pane. After selecting the physical property the designer is prompted to name the new estimation procedure. The names of all existing estimation procedures are displayed in the Procedures Pane.

The physical property and procedure name are used to create an estimation procedure graph object. This new object is added to the Procedure Development Pane.

**Deselect all Procedures:** This command deselects all the estimation procedures displayed in the Procedures Pane. Selection is used by the **Save Selected Procedures** command.

**Document Procedure:** Displays the procedure's documentation in a window resource exposed over the Procedure Development Pane. The command prompts the designer for an estimation procedure. All estimation procedures known to the system

are displayed in the Procedures Pane. The documentation of an estimation procedure can be modified using the **Edit Procedure Documentation** command.

**Document Technique:** Displays the techniques's documentation in a window resource exposed over the Procedure Development Pane. The command prompts the designer for an estimation technique. Any of the estimation techniques displayed in the Procedure Development Pane are available for selection.

**Edit Procedure Documentation:** Places the documentation for a chosen estimation procedure in the Documentation Editing Pane. This pane is a Zwei editor and thus provides complete editing facilities. The command first prompts the designer for an estimation procedure. Termination of editing is done by pressing the <End> key. All carriage returns are removed from the entered documentation. The resulting string is stored in the estimation procedure object.

**Redisplay Development Pane:** Redisplays all the objects of the Procedure Development Pane.

**Respecify Property:** This command enables the designer to change the specified estimation technique for a particular property. The command prompts the designer for a property node. All estimation techniques and property nodes emanating from this property node are removed. They are replaced with all the estimation techniques applicable to estimate the property.

**Save Selected Procedures:** Saves each of the selected procedures of the Procedures Pane into the file:

```
molecular-design:evaluation;procedure-instance.lisp.
```

**Specify Value:** Occasionally it is desirable to specify a value for a physical property. This is useful when investigating a new estimation technique. This command first prompts the designer for a physical property. Next the designer is prompted for a value for this physical property. The value is displayed as the child of the physical property graphically being connected by an arrow.

## 10.6 Values Configuration Commands

The following commands are listed in the Command Menu of the Evaluation Section.

The definitions of these commands are in the file:

```
molecular-design:evaluation;commands.lisp.
```

**Clear Estimated Values Pane:** Removes all the molecules and their estimated properties from the Estimated Values Pane.

**Deselect all Molecules:** Deselects all the molecules displayed in the Estimated Values Pane. Selection is used by the **Estimate Property** command.

**Estimate Property:** Estimates values for a chosen physical property for each of the selected molecules of the Estimated Values Pane. The command first prompts the

designer for a physical property to be estimated. The physical properties known to the system are displayed in a window resource exposed over the Groups Pane. Values for any required state variables are prompted for. The chosen physical property is then estimated for each selected molecule displayed in the Estimated Values Pane. The estimated values are displayed for each molecule.

**Group Enter Molecule:** The estimation procedures currently used in the system all use group contribution techniques to estimate physical properties directly from molecular structure. Entering a molecule for estimation thus consists of entering its groups. This command prompts the designer for a sequence of groups. These groups are stored in an estimated molecule object and displayed in the Estimated Values Pane.

**Menu Enter Molecule:** The interactive design section's **Save Current Molecule** command creates an estimated molecule object. This molecule is placed on a list accessible to the rest of the system. The **Menu Enter Molecule** command allows the designer to select molecules from this list for display in the Estimated Values Pane thus allowing for property estimation.

**Plot Dependent Property:** Physical properties which are dependent upon state variables are able to be plotted using the graph facilities of the Data Base Section. The command prompts the designer for an estimated physical property displayed in the Estimated Values Pane. The configuration is then changed to the Plot Configuration of the Data Base Section. A plot is constructed of the physical property versus its state variables values. If the physical property is dependent upon more than one state

variable the designer is prompted to choose one.

Figure 10.7 shows the Estimated Values Pane with a number of molecules displayed. The first molecule in the pane, dichlorodifluoromethane, has had its  $P_{vp}$  estimated at six temperatures. Choosing the **Plot Dependent Property** command the system changes to the Plot Configuration of the Data Base Section and creates a plot of  $P_{vp}$  vs.  $T$ . This plot is shown in Figure 10.8.

**Select all Molecules:** Selects all the molecules displayed in the Estimated Values Pane. Selection is used by the **Estimate Property** command.

## 10.7 Section Discussion

The “decision tree” approach the Evaluation Section uses to construct estimation procedures is clear and simple. I believe that this approach should be considered for other systems which are investigating methods for organized, flexible construction of equations. The Target Transformation Section should be reimplemented along these lines.

Completing the implementation of “default estimation procedures” would enable the system to suggest portions of new estimation procedures. This would be very useful for complex estimation procedures. Such a “default” system would be extremely applicable to the Target Transformation Section. This possibility is discussed further in Chapter 7.

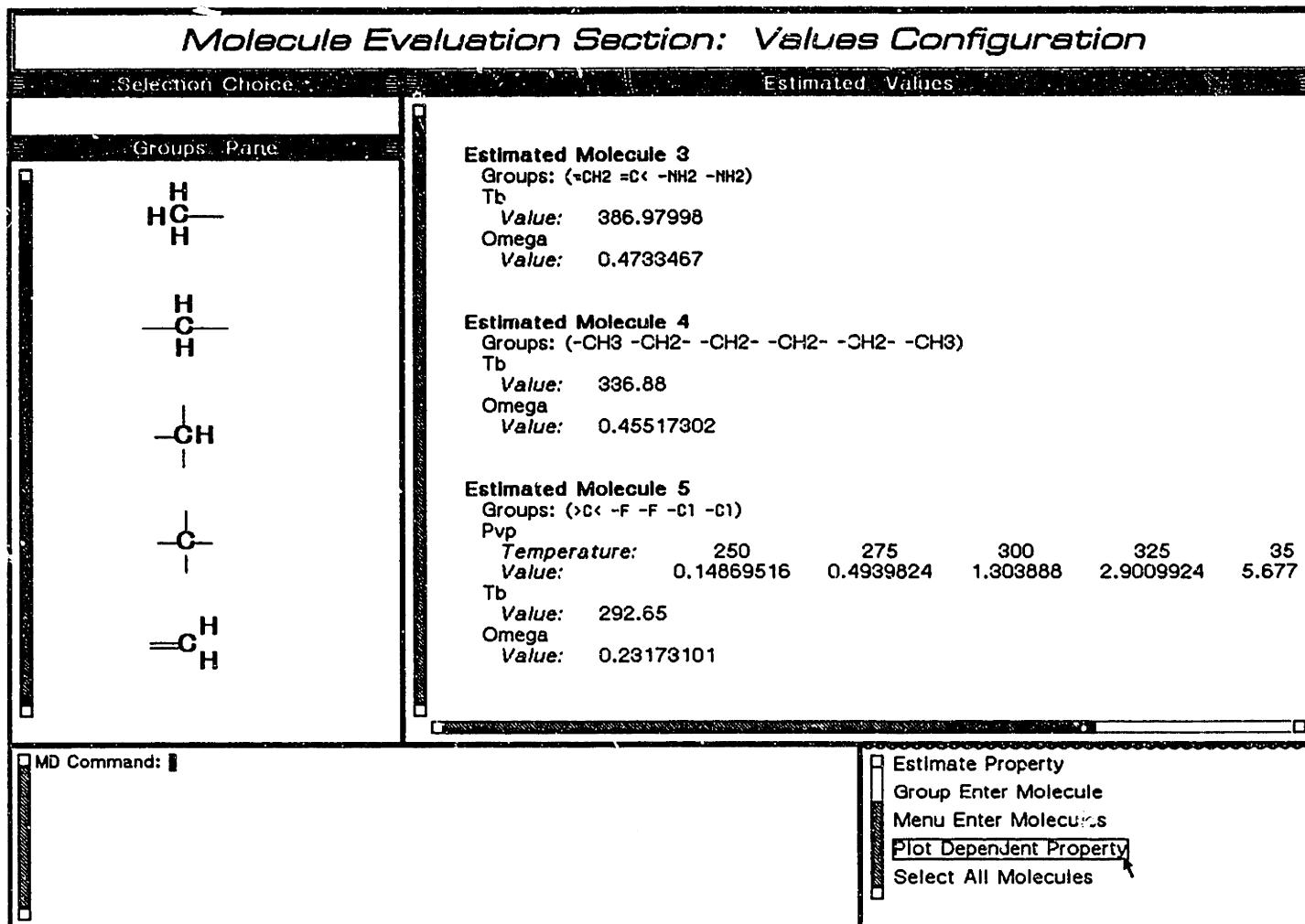
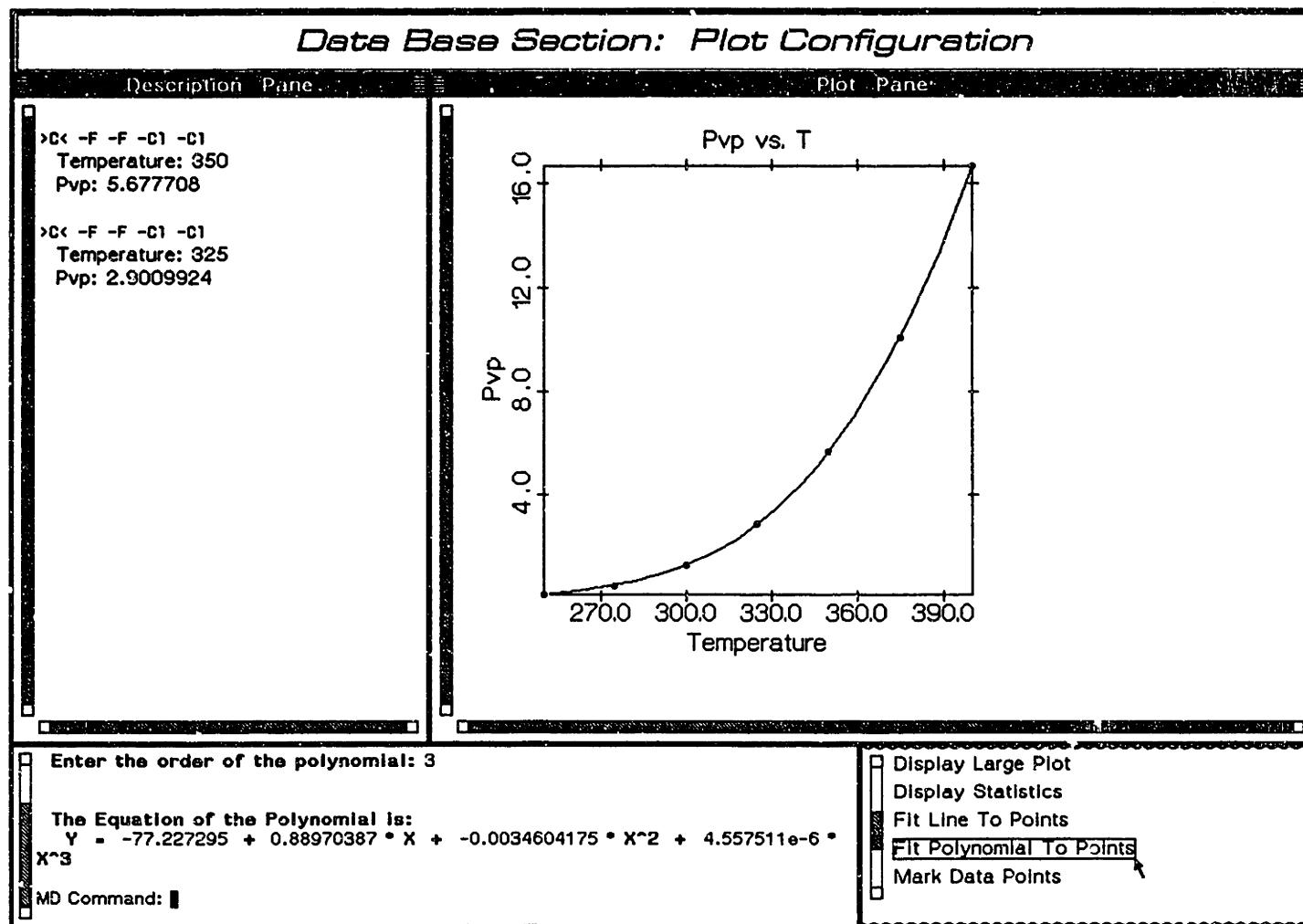


Figure 10.7: Values Configuration with Estimated Physical Properties

Figure 10.8: Estimated  $P_{vp}$  vs.  $T$  Plot

## 10.8 Section Example

These instructions detail a step-by-step example usage of the Evaluation Section. In this example we:

1. Create a new estimation procedure.
2. Enter molecules for estimation.
3. Estimate physical properties.

### Move to the Specifications Configuration

If the current configuration is not the Evaluation Section's Specifications Configuration the first task is to change configurations.

**Action 10.1** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 10.2** *Mouse left on the Molecule Evaluation Section Specifications Configuration.*

The system changes configuration to the Specifications Configuration.

### Creating an Estimation Procedure

The Specifications Configuration provides facilities for developing estimation procedures. We develop an estimation procedure for estimating the vapor pressure.

**Action 10.3** *Mouse left on the Create Estimation Procedure command.*

The system first prompts for the property to be estimated:

**Enter the Property to be Estimated:**

The physical properties known to the system are displayed in a window resource exposed over the Procedures Pane. The physical properties are arranged by property class.

**Action 10.4** *Mouse left on the Vapor Pressure property displayed in the exposed window resource.*

The system keeps track of estimation procedures using names specified by the designer.

The Procedures Pane lists the names of all estimation procedures known to the system.

The system prompts for a name for our new procedure:

**Enter a Name for this New Procedure:**

**Action 10.5** *Type into the Molecular Design Interaction Pane: Pvp New. Press return when the entry is complete.*

The system displays the initial tree structure consisting of the property being estimated,  $P_{vp}$ , and the estimation techniques known to the system which estimate it. The display of the system at this point should be similar to Figure 10.3.

We now choose estimation techniques:

**Action 10.6** *Mouse left on the Choose Technique command.*

The system prompts for an estimation-technique-node:

**Enter an Estimation Technique Node:**

**Action 10.7** *Mouse left on the estimation technique node:*

$P_{-p} \rightarrow (T_b \ T_{br} \ P_c)$  - Riedel Plank Miller Tbr Biased

*displayed in the Procedure Development Pane.*

The chosen estimation technique is “assigned” to the physical property. The physical property display is changed to display the physical property with the chosen technique displayed beneath it. The required properties of the estimation technique are displayed as the children of this new node. Figure 10.9 shows the appearance of the system after these actions.

We continue to choose estimation techniques:

**Action 10.8** *Mouse left on the Choose Technique command.*

The system prompts for an estimation-technique-node:

**Enter an Estimation Technique Node:**

**Action 10.9** *Mouse left on the estimation technique node:*

$T_b \rightarrow ()$  - Joback Tb

*displayed in the Procedure Development Pane.*

**Action 10.10** *Mouse left on the Choose Technique command.*

The system prompts for an estimation-technique-node:

**Enter an Estimation Technique Node:**

## Molecule Evaluation Section: Specifications Configuration

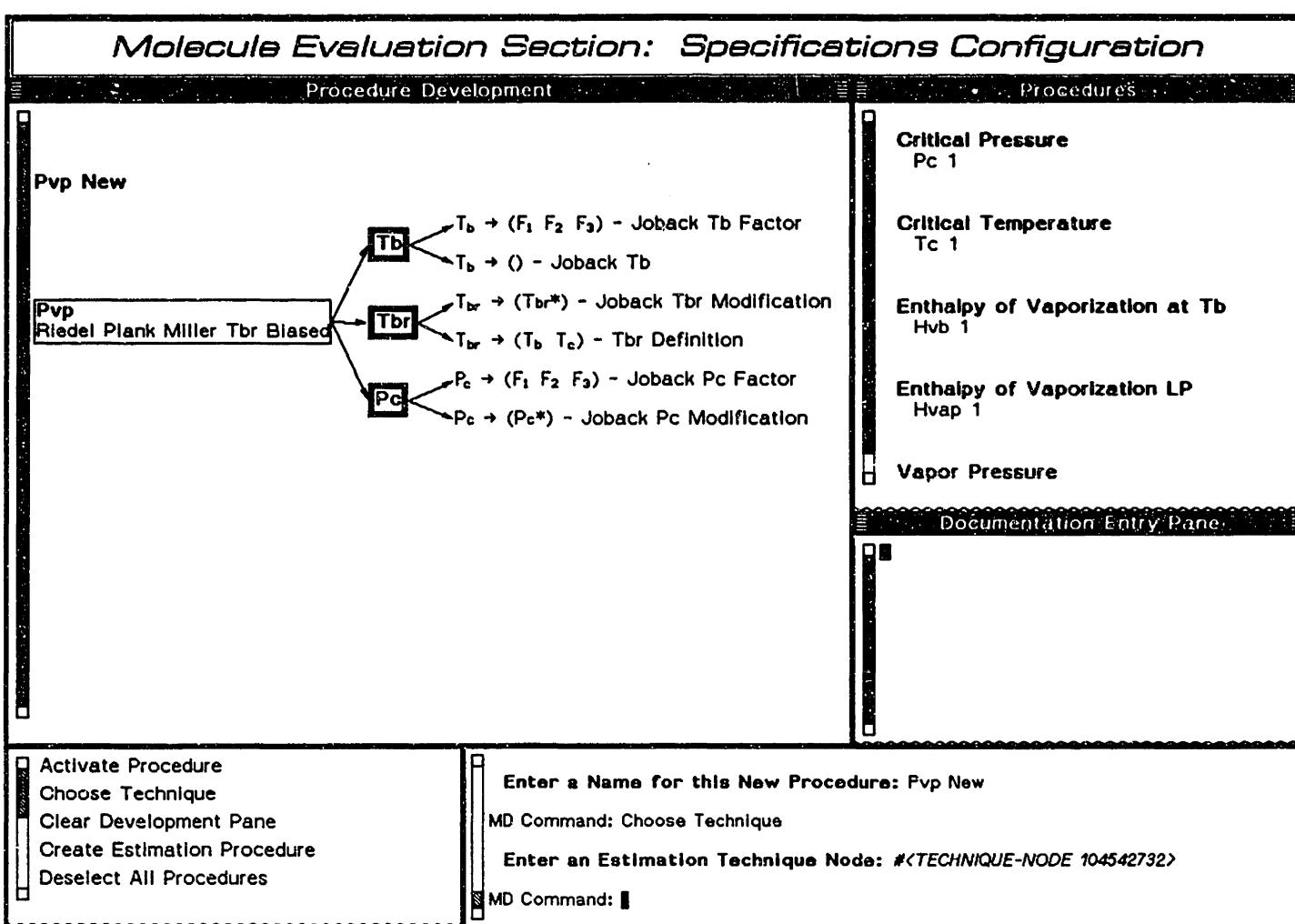


Figure 10.9: Estimation Technique Choices after  $P_{vp}$  Technique Selection

**Action 10.11** *Mouse left on the estimation technique node:*

$T_{br} \rightarrow (T_{br}^*)$  - Joback *Tbr Modification*

*displayed in the Procedure Development Pane.*

**Action 10.12** *Mouse left on the Choose Technique command.*

The system prompts for an **estimation-technique-node**:

**Enter an Estimation Technique Node:**

**Action 10.13** *Mouse left on the estimation technique node:*

$P_c \rightarrow (P_c^*)$  - Joback *Pc Modification*

*displayed in the Procedure Development Pane.*

**Action 10.14** *Mouse left on the Choose Technique command.*

The system prompts for an **estimation-technique-node**:

**Enter an Estimation Technique Node:**

**Action 10.15** *Mouse left on the estimation technique node:*

$T_{br}^* \rightarrow ()$  - Joback *Tbr\**

*displayed in the Procedure Development Pane.*

**Action 10.16** *Mouse left on the Choose Technique command.*

The system prompts for an **estimation-technique-node**:

**Enter an Estimation Technique Node:**

**Action 10.17** *Mouse left on the estimation technique node:*

$P_c^* \rightarrow () - Joback P_c^*$

*displayed in the Procedure Development Pane.*

All the estimation techniques are now chosen. The system display resembles Figure 10.5. The technique specifications displayed in the New Pvp estimation procedure graph are now used to form the LISP code implementing the procedure.

**Action 10.18** *Mouse left on the Activate Procedure command.*

The system prompts for an **estimation-procedure-graph**:

**Enter an Estimation Procedure Graph:**

**Action 10.19** *Mouse left on the Pvp New estimation procedure graph displayed in the Procedure Development Pane.*

The system constructs the LISP function and adds the name of our new estimation procedure to the Procedures Pane.

## **Entering Molecules**

The Values Configuration provides facilities for using the developed estimation procedures. The first task is to move to the Values Configuration.

**Action 10.20** *Mouse right on an empty area of the screen.*

A menu is exposed containing all the configurations of the system arranged by section.

**Action 10.21** *Mouse left on the Evaluation Section Configuration.*

The system changes configuration to the Data Configuration.

We enter the molecules whose properties are being estimated.

**Action 10.22** *Mouse left on the Group Enter Molecule command.*

The system prompts for a sequence of groups:

**Enter a Sequence of Groups:**

We enter the five groups which make up 1,1,2-trifluoroethane.

**Action 10.23** *Mouse left on the >CH- group displayed in the Groups Pane.*

**Action 10.24** *Mouse left on the -CH2- group displayed in the Groups Pane.*

**Action 10.25** *Mouse left on the -F group displayed in the Groups Pane.*

**Action 10.26** *Mouse left on the -F group displayed in the Groups Pane.*

**Action 10.27** *Mouse left on the -F group displayed in the Groups Pane.*

**Action 10.28** *Once these groups are entered press the return key.*

The system collects the entered groups and forms an **estimated-molecule**. This molecule is named “Estimated Molecule 1” and added to the Estimated Values Pane.

We enter two more molecules:

**Action 10.29** *Mouse left on the Group Enter Molecule command.*

The system prompts for a sequence of groups:

**Enter a Sequence of Groups:**

We enter the five groups which make up dichlorodifluoromethane.

**Action 10.30** *Mouse left on the >C< group displayed in the Groups Pane.*

**Action 10.31** *Mouse left on the -Cl group displayed in the Groups Pane.*

**Action 10.32** *Mouse left on the -Cl group displayed in the Groups Pane.*

**Action 10.33** *Mouse left on the -F group displayed in the Groups Pane.*

**Action 10.34** *Mouse left on the -F group displayed in the Groups Pane.*

**Action 10.35** *Once these groups are entered press the return key.*

The system collects the entered groups and forms an **estimated-molecule**. This molecule is named “Estimated Molecule 2” and added to the Estimated Values Pane.

We enter one more molecule:

**Action 10.36** *Mouse left on the Group Enter Molecule command.*

The system prompts for a sequence of groups:

**Enter a Sequence of Groups:**

We enter the seven groups which make up heptane.

**Action 10.37** *Mouse left on the -CH<sub>3</sub> group displayed in the Groups Pane.*

**Action 10.38** *Mouse left on the -CH<sub>2</sub>- group displayed in the Groups Pane.*

**Action 10.39** *Mouse left on the -CH2- group displayed in the Groups Pane.*

**Action 10.40** *Mouse left on the -CH2- group displayed in the Groups Pane.*

**Action 10.41** *Mouse left on the -CH2- group displayed in the Groups Pane.*

**Action 10.42** *Mouse left on the -CH3 group displayed in the Groups Pane.*

**Action 10.44** *Once these groups are entered press the return key.*

The system collects the entered groups and forms an **estimated-molecule**. This molecule is named “Estimated Molecule 3” and added to the Estimated Values Pane.

## **Estimating Properties**

The system estimates the properties for all selected molecules displayed in the Estimated Values Pane.

**Action 10.45** *Mouse left on the Select All Molecules command.*

**Action 10.46** *Mouse left on the Estimate Property command.*

The system prompts for the property to be estimated:

**Enter the property to be estimated:**

The physical properties known to the system are displayed in a window resource exposed over the Groups Pane.

**Action 10.47** *Mouse left on the Vapor Pressure physical property displayed in the window resource.*

Since the vapor pressure is temperature dependent, the system prompts for a sequence of temperatures:

**Enter a sequence of Temperature Values:**

**Action 10.48** *Type in the following sequence of temperatures:*

300      325      350      375      400

*Separate the entries by commas. Do not put a space after the comma; the system does this for you.*

**Action 10.49** *Once the sequence of temperatures is entered press the return key.*

The system prompts for an estimation procedure to use to estimate the vapor pressure:

**Enter an Estimation Procedure:**

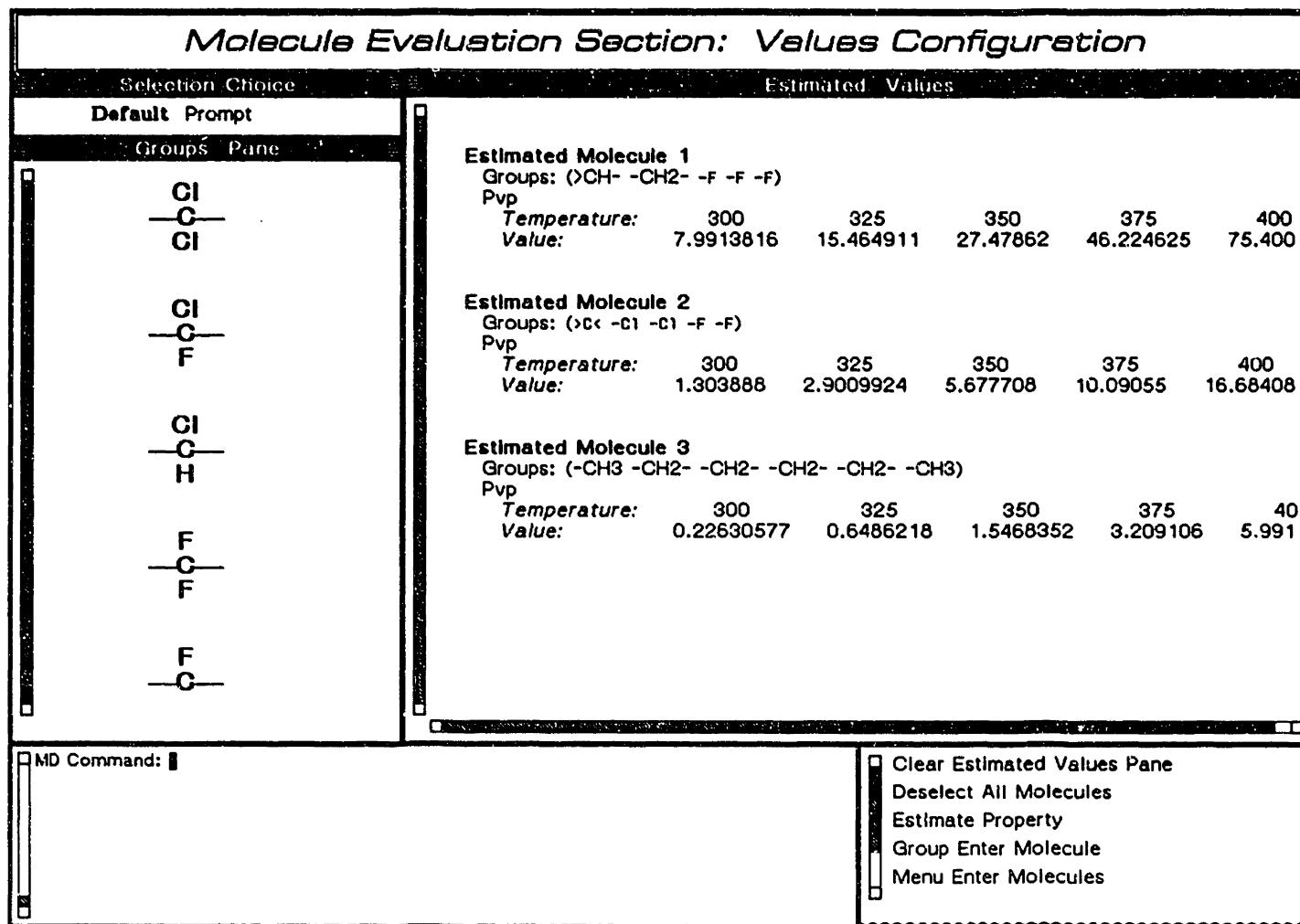
Applicable estimation procedures are displayed in a window resource exposed over the Groups Pane.

**Action 10.50** *Mouse left on the Pvp New estimation procedure.*

The system estimates the vapor pressure at five specified temperatures for each of the estimated molecules displayed in the Estimated Values Pane.

**Action 10.51** *Mouse left on the Deselect All Molecules command.*

Figure 10.10 shows the screen display.

Figure 10.10: Estimating  $P_{vp}$  Values for Three Molecules

## **Plotting Estimates**

The system provides the facility for plotting temperature or pressure dependent physical properties.

**Action 10.52** *Mouse left on the Plot Dependent Property command.*

The system prompts for an estimated physical property:

**Enter a Dependent Estimated Property:**

**Action 10.53** *Mouse left on the estimated  $P_{vp}$  values displayed for Estimated Molecule 1.*

The system changes to the Data Base Section's Plot Configuration. The estimated  $P_{vp}$  values are plotted as a function of temperature.

# Chapter 11

## Final Remarks

The previous chapters describe the implementation of my molecular design system. Each chapter describes a section of the implementation. In this final chapter I present some conclusions and some general comments on possible future directions.

### 11.1 System Status

Despite several limitations identified in previous chapters, the system stands complete in its ability to design molecules via my interactive and automatic design procedures. Supporting facilities for displaying data, estimating physical properties, creating new groups, and entering new estimation techniques are available and functional.

### 11.2 Extending the System

The system can be extended in two directions: 1) Entering additional physical properties and estimation techniques enables the system to design more classes of chemical

products; 2) Improving the underlying algorithms – increasing the speed of the automatic design's search procedure and the interactive design's solution procedure.

The structure of many physical property estimation techniques preclude their use in the current algorithms. UNIFAC and molecular modeling techniques are two examples. Investigating such new techniques will lead to fundamental changes in the algorithms used or entirely new algorithms.

Improving the efficiency of the automatic design algorithm would permit performing numerous case studies. These case studies could concentrate on evaluating division and expansion strategies.

My recommendation is to use the design of new classes of physical properties to improve and add to the existing algorithms. Support facilities for developing estimation techniques need to be extended. Tools for analyzing estimation techniques need to be added.

### **11.3 LISP**

LISP is an amazing language. My computer experience was typical of a chemical engineer. In freshman year I took a half semester course on FORTRAN programming. After writing numerous programs throughout my undergraduate days and conducting my master's thesis research entirely in FORTRAN I had a very good command of the language. Coming from a FORTRAN background the idea of implementing large programs which were easily modifiable, incrementally changeable, consisting of numerous data structures, and able to perform a multitude of complex tasks seemed impossible.

On exposure to LISP suddenly these and more tasks not only seemed possible but a natural outgrowth of the language itself.

I feel a great disservice has been paid to the large number of people who constantly yearn to understand the new developments in AI and computer science. The concepts of modularity, ease of modification, incremental development, and automatic code generation which are extolled by expert systems and object oriented languages are portrayed as revolutionary concepts. They are revolutionary only to those individuals who have not programmed in LISP.

Recommendation: LEARN LISP.

## 11.4 Persistence

One of the major problems with the construction of complex data structures is the inability to store these structures in files for future use. Typical file system operations are inadequate to provide the capabilities to edit and store structures easily. Unfortunately the remedy for this problem is typically to simplify the data structure.

Many of the objects used to represent properties, estimation techniques, and groups were designed to facilitate file storage. However, once stored these objects can only be further manipulated by editing the file they are stored in with a conventional text editor.

Object oriented data-bases provide an excellent solution to the problem of object storage. The data-bases provide the same powerful representation capabilities as flavors but are designed to allow data storage with no additional implementation by the user.

I recommend object oriented data-bases be considered for any future development.

# Bibliography

- [1] Ambrose, D.: *Correlation and Estimation of Vapor-Liquid Critical Properties: II. Critical Pressures and Volumes of Organic Compounds*. National Physical Laboratories Report Chem 98, 1979.
- [2] Ambrose, D.: *Correlation and Estimation of Vapor-Liquid Critical Properties: I. Critical Temperatures of Organic Compounds*. National Physical Laboratories Report Chem 92, 1980.
- [3] Apple Computer Company: *Inside Macintosh*. Volume 1, Addison-Wesley, Reading Massachusetts, 1985.
- [4] Bromley, Hank: *LISP Lore: A Guide to Programming the LISP Machine*. Kluwer Academic, Boston, 1986.
- [5] Goodman, Danny: *The Complete HyperCard Handbook*. Bantam Books, Toronto, 1987.
- [6] Gray, Neil A. B.: *Computer-Assisted Structure Elucidation*. John Wiley & Sons, New York, 1896.
- [7] Keene, Sonya E.: *Object-Oriented Programming in Common LISP*. Addison-Wesley, Reading Massachusetts, 1989.
- [8] Lynch, M. F., J. M. Harrison, W. G. Town, and J. E. Ash: *Computer Handling of Chemical Structure Information*. American Elsevier, New York, 1971.
- [9] McClellan, A. L.: *Tables of Experimental Dipole Moments*. Freeman, San Francisco, 1963.

- [10] National Academy of Sciences. *Survey of Chemical Notation Systems*. Publication Number 1150. Washington, D.C., 1964.
- [11] National Academy of Sciences. *Survey of European Non-Conventional Chemical Notation Systems*. Publication Number 1278. Washington, D.C., 1965.
- [12] National Academy of Sciences. *Survey of Chemical Information Handling*. Publication Number 1733. National Academy of Sciences, Washington, D.C., 1969.
- [13] Orrick, C.: *Estimation of Viscosity for Organic Liquids*. July 25, 1973.
- [14] Reid, R. C., J. M. Prausnitz, and T. K. Sherwood: *The Properties of Gases and Liquids*. McGraw-Hill Book Company, New York, 1977.
- [15] Smith, E. G.: *The Wiswesser Line-Formula Chemical Notation*. McGraw-Hill, New York, 1968.
- [16] Steele Jr., Guy L.: *Common LISP*. Digital Press, Bedford, Massachusetts, 1984.
- [17] Stull, D. R., E. F. Westrum, and G. C. Sinke: *Chemical Thermodynamics of Organic Compounds*. John Wiley and Sons, Inc., New York, 1969.
- [18] Symbolics Reference Manuals. Symbolics, Cambridge, Massachusetts, 1988.
- [19] Winston, Patrick Henry and Berthold Klaus Paul Horn: *LISP*. Addison-Wesley, Reading, Massachusetts, 1984.

# Appendix A

## Linear Names

Linear codes are extremely important for entering molecular structures into the computer. Wiswesser Line Notation(WLN)[15] is the most widely used notation. The Smiles Code is now gaining popularity. Besides the ability to enter molecular structure information via non graphics terminals, the speed and accuracy of entry is improved.

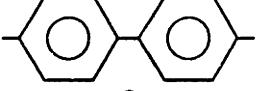
I have simply named each of the groups used in the system. The majority of the names are obtained directly from their molecular formulas. Some groups have names not derivable from their molecular formulas. Table A.1 shows these names. Adaptation of either Wiswesser Line Notation or the Smiles Code or strict adherence to the IUPAC naming convention is recommended for future work.

I constructed a character style with five new characters:

$\overline{r}$     $\overset{r}{>}$     $\overset{r}{>}$     $\overset{r}{<}$     $\overset{r}{<}$

The character style is named Group.Roman.Normal. Since there are no keys for such characters on the Symbolics keyboard it was necessary to reassign certain keys. I

Table A.1: Complex Groups' Linear Names

	Group	Name
1)		pphenyl
2)	$-\text{CH}_2-\text{C}_6\text{H}_4-$	$-\text{CH}_2\text{-pphenyl}$
3)	$-\text{CH}_2-\text{C}_6\text{H}_4-\text{CH}_2-$	$-\text{CH}_2\text{-pphenyl-CH}_2\text{-}$
4)	$-\text{C}_6\text{H}_4-\text{CH}_2-\text{C}_6\text{H}_4-$	$-\text{pphenyl-CH}_2\text{-pphenyl-}$
5)		$-\text{pphenyl-pphenyl-}$
6)	$-\text{O}-\text{C}=\text{O}-$	$-\text{O}-\text{COO}-$

chose to reassign some of the **symbol** keys. These are not frequently used. Thus when the current character style is **Group.Roman.Normal** typing a **sh-sy-p** produces the character  $\text{ $\pi$ }_r$  instead of the character  $\pi$ . In all other character styles and in hardcopy the character will be displayed as  $\pi$ . Table A.2 shows the list of redefined symbol

Table A.2: Bond Symbols

Bond	Symbol	Key Sequence <sup>†</sup>
$\text{\overline{r}}_r$	$\beta$	<b>sh-sy-b</b>
$\text{>}_r$	$\bullet$	<b>sy-'</b>
$\text{>}_r$	$\wedge$	<b>sy-q</b>
$\text{<}_r$	$\epsilon$	<b>sh-sy-e</b>
$\text{<}_r$	$\pi$	<b>sh-sy-p</b>
$\equiv$	$\equiv$	<b>sh-sy-'</b>

<sup>†</sup> **sh** = the shift key; **sy** = the symbol key.

characters and keystrokes for the six new “bond” characters.

# Appendix B

## Preparation

Understanding the implementation of my molecular design system requires understanding the LISP language and the Symbolics environment. My system is implemented in Symbolics Common LISP(SCL). SCL is a superset of Common LISP containing many extensions developed by Symbolics. In this appendix I present an outline for learning SCL and the Symbolics development environment.

One word of advice. LISP is a language of great power and subtlety. A handful of primitives enable you to program impressive applications. Over time you will find that LISP's explicit scoping, ability to treat code as data, and simple syntax result in an extremely well organized language and code. Appreciating this organization will enable you to produce applications of immense power. However, to obtain this appreciation you must continually strive to improve your programming. This means reexamining your code for improved ways of implementation. Your manager or advisor will probably see this as a waste of time. "If it works, why fix it?" However, remember that as you continue to build your structure ever higher you had better check the

foundation to ensure it can handle the additional burden. To those who berate you for being unproductive, simply ask them how many thousands of lines of LISP code have they written.

## B.1 References

I refer to five references in this appendix:

1. *Common LISP* by Guy L. Steele Jr.[16]
2. *LISP* by Patrick Winston and Berthold Horn[19]
3. Symbolics Reference Manuals by Symbolics Incorporated[18]. The reference set I refer to is for the Genera 7.1 release.
4. *LISP Lore* by Hank Bromley[4].
5. Object-Oriented Programming in Common LISP by Sonya Keene[7].

Each of these references has a different emphasis. Steele is an invaluable reference for the Common LISP language. However, its discussion of the language is too terse for a first book on learning LISP. Winston and Horn provides a good introduction to Common LISP. The text is well organized presenting many Common LISP primitives and a great deal of example code. The Symbolics Reference Manuals describe the Symbolics environment adequately. Volume 2A, describing Common LISP and SCL primitives, and Volume 7A, describing interface programming, are the two volumes most needed for typical programming tasks. Many examples are provided but the emphasis is on reference. Bromley presents more of an instructional approach to learning the LISP machine. Keene describes the new Common LISP Object System(CLOS) standard. Although different from the Flavors system used by Symbolics it provides a nice introduction to object oriented programming concepts.

## B.2 Background

Readings:

- LISP: Chapter 1
- Common LISP: Chapter 1

These chapters introduce LISP's style as a programming language. Winston and Horn's Chapter 1 provides some ammunition against opponents of LISP.

## B.3 Data Types

Readings:

- Common LISP: Chapter 2
- Symbolics: Volume 2A, Chapters 1 and 5
- Common LISP: Chapter 4

LISP provides many data types. For many applications the only data types you need to be concerned about are numbers and lists. Later you will learn about structures and flavors which are means for generating new data types. Unlike other languages you do not have to declare variables to be of a certain type in LISP. Data types are thus best used as a means for organizing your thoughts about information storage and knowledge representation.

Steele's Chapter 4 should be looked through only briefly. The type specification of common lisp was extended by Symbolics into its presentation system. The presentation system is discussed in Section B.11.

## B.4 Editor

Readings:

- Symbolics: Volume 3, Chapters 1-8

The majority of time during program development is spent in the editor. Understanding the editor is thus important to improving programming productivity. The ZMACS editor on the Symbolics is an extremely powerful editor. Volume 3 of the Symbolics reference manuals describes the editor in detail. The first eight chapters present many editor commands. These should be read lightly. As you continue to program on the Symbolics you should occasionally glance through these chapters to identify any editor commands which could speed your programming.

Symbolics provides an online ZMACS tutorial. The tutorial covers only the basic ZMACS commands but is a very good place to start. The file:

**sys:examples;Teach-Zmacs-Info.text**

tells you how to run the tutorial. To view the file enter:

**Show File sys:examples;Teach-Zmacs-Info.text**

## B.5 Basics

Readings:

- LISP: Chapters 2, 3, 4, 5
- Common LISP: Chapters 3, 5, 6, 7
- Symbolics: Chapter 21.8

Problems:

- LISP: 3-5, 3-6, 4-10, 4-20, 4-29

Winston and Horn's Chapter 4 discusses *Recursion and Iteration*. SCL provides the `loop` macro as a powerful facility for performing iteration. The `loop` macro is not part of common lisp and is thus not discussed in Winston and Horn or Steele. An extensive discussion of the `loop` macro is found in Chapter 21.8 of the Symbolics Reference Volume 2A, Genera 7.1 edition. I recommend doing the examples shown in Chapter 21.8 and doing the examples of Winston and Horn's Chapter 4 using the `loop` macro. Almost anything done using the LISP primitive `do` is done easier and cleaner with `loop`. Problem 4-20 can be done using the Symbolics `graphics:draw-line` function for the `line` function. The function call would look like:

```
(graphics:draw-line x-start y-start x-end y-end)
```

Winston and Horn's Chapter 5 briefly mentions *structures*. Structures are useful for organizing simple information and provide a good introduction to the Flavors system. Steele's Chapter 19 provides a more extensive discussion of structures.

## B.6 Macros

Readings:

- Common LISP: Chapter 8
- LISP: Chapter 8
- Symbolics: Chapters 13, 14, 15

Macros enable the programmer to write code which writes code. Macros make it possible to write code that is clear and elegant at the user level, but is converted to a

more complex or more efficient internal form for execution. Chapter 15.6 of Symbolics' reference manuals discusses some helpful hints for writing macros.

## B.7 Debugger

Readings:

- Symbolics: Volume 4, Chapters 1, 2, 3, 4

Symbolics provides a very powerful debugger. The readings should be done very lightly.

It is best to try using the debugger when you get an error during programming. Four of the most useful debugging functions I have found are:

1. **c-N**: moves to the next frame.
2. **c-P**: moves to the previous frame.
3. **(setf (dbg:arg *n*) *x*)**: sets the value of the *n*th argument in the current frame to the value of *x*.
4. **c-m-R**: restarts execution of the function in the current frame.
5. **c-E**: edits the current function.

These functions can be used to move between function calls, change the values of function arguments, and reinvckle the function call.

## B.8 I/O

Readings:

- Common LISP: Chapters 21, 22, 23
- LISP Lore: Chapter 6
- Symbolics: Volume 5

Reading and writing to the screen or files is facilitated by the use of streams. The concept of streams is introduced in chapter 21 of Common LISP. The extensive array of I/O functions described in chapters 22 and 23 demonstrate the use of streams. Chapter 6 of LISP Lore and Symbolics' Volume 5 describe the use of streams on the LISP machine.

One of the most common types of streams on the LISP machine are windows. Windows are implemented as flavors on the Symbolics. They are thus discussed in Section B.10 after flavors are described.

Symbolics provides many extensions of I/O facilities in its interface facilities.

## B.9 Flavors

Readings:

- OOP: Chapters 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- LISP Lore: Chapters 2
- Symbolics: Volume 2A, Chapters 3.2, 17

Flavors are a major component of all programming done on the Symbolics. Understanding them is important. Although Object Oriented Programming in Common LISP discusses another object oriented language, CLOS, many of the concepts are applicable to programming in Flavors. Chapter 3.2 of Symbolics' Volume 2A provides a good introduction to the basic concepts, usefulness, and syntax the Flavors. The ship example used in chapter 2 of LISP Lore should be worked through.

## B.10 Windows

Readings:

- LISP Lore: Chapter 5

Problems:

- LISP Lore: Chapter 2; Problems 1, 2, 3, 4, 5, 6

Windows combine your knowledge of Flavors and streams. The references above give an overview of windows on the Symbolics. Possibly the best way to become familiar with windows is to begin trying them. The function `tv:make-window` is the basic facility for creating windows.

Windows are often used in aggregates called frames. Symbolics provides the Frame-Up activity for the construction of program frames.

## B.11 Presentations

Readings:

- Symbolics: Volume 7A, Chapter 7

The presentation system is one of the newer additions to the Symbolics system. It is immensely powerful. The documentation provided in Volume 7A is not adequate. The documentation for Genera 7.2 discusses presentations more clearly and in more detail.

## B.12 Interface Programming

Readings:

- Symbolics: Volume 7A, All

Programs in general perform three tasks: 1) accept input; 2) process information; 3) present results. The interface between user and program thus participates in two-thirds of all programming tasks. Symbolics has developed a paradigm for interface development and numerous facilities for implementing this paradigm. Volume 7A describes this paradigm and the programming facilities.

The example code described in chapter 9 is in the file:

**sys:examples;ui-application-example.lisp.**

I suggest starting with this example. Particular attention should be paid to the use of:

1. **program-frames**
2. **presentation-to-command-translators**
3. **formatting macros**
4. **command usage**

## **B.13 Proceeding**

Symbolics provides source code for most of its system. The majority of this code is very well written and serves as an excellent example of how to write LISP on the Symbolics.

Whenever you see a system behavior you would like to use, track down the source code, copy it to another file and modify to your needs.

# Appendix C

## Installation

My molecular design system is distributed on a cartridge tape. The tape contains both the source and binary files. The molecular design system was constructed using Symbolics' System Construction Tool facility. Loading and activation of the system is done analogously to other layered products you run on the LISP machine.

### C.1 System Requirements

The system has two requirements:

1. The system runs under Genera 7.1. Attempting to run the system under 7.2 causes an irrecoverable system error.
2. The system's source and binary files require 782 LMFS blocks. The source files by require 420 blocks. The binary files require 358 blocks.

### C.2 System Restoration

The system is loaded onto a cartridge tape in *Distribution Format*. This is the same format which Symbolics uses to distribute its layered products.

Before restoration begins it is necessary to create the translations file for the system.

The name of this file must be:

**molecular-design.translations**

The translations file defines the logical host:

**molecular-design**

which is used throughout the system. The translations file specifies the location of the system files.

The translations file is located in your **sys:site;** directory. The following is an example translations file in which the system files are to be stored in the directory

```
Fungus:>kevin>final-thesis>

;;; -*- Syntax: Common-Lisp; Package: User; Mode: LISP -*-
(fs:set-logical-pathname-host "molecular-design"
  :physical-host "Fungus"
  :translations '(("molecular-design:**;*.*.*"
    "Fungus:>kevin>final-thesis>**>*.*.*")))
```

To successfully restore the system from tape you must first load the translations file. Restoration of the files from tape to disk is done by placing the tape into the tape drive and issuing the follow commands:

**Load File sys:site;molecular-design.translations**

**Restore Distribution**

When the menu pops up click on **Do It**. The 92 files will be restored from tape to the directory specified in your translations file.

## C.3 Loading

My molecular design system is constructed using Symbolics's *System Construction Tool*. The various commands which are applicable to systems are thus applicable to the molecular design system.

To load the system into the LISP world use the following command:

```
Compile System Molecular-Design :Redefinitions Ok Yes
```

This command loads 8 font files and 40 compiled lisp files. The machine displays a message as it loads each file. Loading all the files takes about 17 minutes.

The `:Redefinition Ok` keyword is specified to allows the system's `margin-component-draw` and `:highlight-string-intervals` generic functions to be redefined.

Once the files have been successfully loaded the system is activated by typing `<select> ●`. Here `●` denotes the top center key on the Symbolics keyboard displaying a filled in circle. The system will take approximately 2 minutes to activate depending on the particular machine configuration.