Supporting Information For
Rapid Optical Determination of Enantiomeric Excess, Diastereomeric Excess, and Total Concentration Using Dynamic-Covalent Assemblies. A Demonstration using 2-Aminocyclohexanol and Chemometrics.

Brenden T. Herrera,[†] Sarah R. Moor,[†] Matthew McVeigh,[†] Emily K. Roesner,[†,‡] Federico Marini[§,], and Eric V. Anslyn[†,*]*

[†]Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, USA
[§]Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, Rome I-00185, Italy

Table of Contents

I. General

NMR spectra were recorded from The University of Texas at Austin NMR facility. Circular dichroism (CD) spectra were recorded on either a Jasco J-815 spectropolarimeter in the Targeted Therapeutic Drug Discovery and Development Program facility at the University of Texas at Austin or an EKKO CD plate-reader designed by Hinds Instruments and commercialized by BioLogic Science Instruments. Fluorescence experiments were conducted using a BioTek Cytation3 plate-reader. All reagents were of the best grade commercially available. (*1R,2R*) and (*1S,2S*) 2-aminocyclohexanol hydrochloride were purchased from OxChem; (*1R,2S*) and (*1S,2R*) 2-aminocyclohexanol hydrochloride were purchased from ArkPharm. Dipicolylamine was purchased from TCI; pyridine-2-carboxaldehyde was purchased from Acros; 3-hydroxy-pyridine-2-carboxaldehyde ws purchased from OxChem. 4-(2-chloroethyl)morpholine hydrochloride and CD_3CN were purchased from Millipore Sigma. Zinc (II) triflate and iron (II) triflate were purchased from Strem Chemicals. The 3Å molecular sieves (4 to 8 mesh, Acros) were activated at 170 °C for 24 h in a vacuum oven, cooled to room temperature and stored in a desiccator.

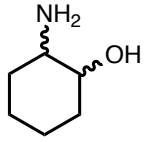
II. Materials and Methods

Synthetic Procedures

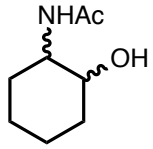
A. Substrate Synthesis

All four stereoisomers [(*1R,2R*), (*1R,2S*), (*1S,2R*), and (*1S,2S*)] of 2-aminocylcohexanol and derivatives thereof were subjected to the same experimental procedures.

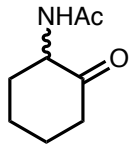
(S1) 2-aminocyclohexanol To a 6-dram vial charged with a stir bar and 2-aminocyclohexanol hydrochloride (3.3 mmol, 1 eq.), was added MeOH (10 mL) and Amberlyst A26 hydroxide resin. The heterogeneous mixture was stirred for 3 hours at room temperature. The resin was removed via vacuum filtration and the filtrate was concentrated under reduced pressure. The resulting white paste was dissolved in $CHCl_3$ and was dried with $MgSO_4$. Solid was removed via filtration, and the filtrate was concentrated under reduced pressure to yield pure product. Characterization of the compound agreed with literature data.[1]



(S2) *N*-(2-hydroxycyclohexyl)acetamide To a 6-dram vial charged with a stir bar and 2-aminocyclohexanol hydrochloride (1 g, 6.6 mmol), was added acetone (7 mL). The suspension was cooled to 0 ºC and 10% Na_2CO_3 (7 mL) was added followed by the slow addition of Ac_2O. The mixture was allowed to warm to room temperature over the course of 3 hours, at which point $NaHCO_3$ (sat. aqueous solution) and solid NaCl were added. The solution was extracted with 9:1 *v:v* $CHCl_3$/isopropanol (5 x 25 mL). The combined organics were dried over Na_2SO_4, filtered and concentrated under reduced pressure to yield pure product in quantitative yield. Characterization of the compound agreed with literature data.[2]



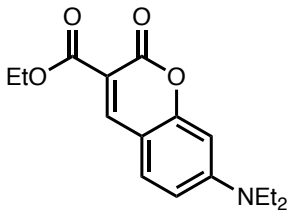
(S3) *N*-(2-oxocyclohexyl)acetamide To a 2-dram vial charged with a stir bar and *N*-(2-hydroxycyclohexyl)acetamide (***2***) (50 mg, 0.318 mmol, 1 eq.), was added CH_2Cl_2 (3 mL). To the stirring solution, was added Dess-Martin periodinane (202 mg, 0.477 mmol, 1.5 eq.) and the mixture was stirred for 3 hours at room temperature. The reaction was quenched with 10 mL of 1:1 *v:v* 10% $Na_2S_2O_3$ (aq.)/sat. $NaHCO_3$ (aq.).



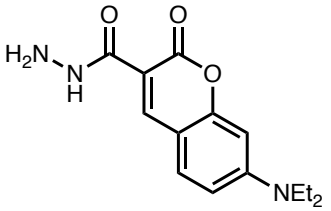
The layers were separated and the organic layer was washed further with sat. $NaHCO_3$ (aq.) (5 X 10 mL), and sat. NaCl (aq.) (1 x 10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated under reduced pressure to yield pure product in quantitative yield. Characterization of the compound agreed with literature data.[3]

B. Hydrazide Fluorophore Synthesis

(S4) Ethyl 7-(diethylamino)coumarin-3-carboxylate To a round bottom charged with a stir bar and 4-diethylaminosalicylaldehyde (955 mg, 5 mmol, 1 eq.), was added anhydrous EtOH (3.3 mL). To this stirring solution was added, diethyl malonate (1 mL, 6.5 mmol, 1.3 eq.) and piperidine (0.4 mL, 2.15 mmol, 0.43 mmol). The mixture was heated to reflux for 2 hours at which point it was cooled to room temperature and quenched with water (5 mL). The mixture was extracted with EtOAc (3 X 25 mL), and the combined organic layer was dried over $MgSO_4$, filtered and concentrated under reduced pressure to yield crude product. The crude product was purified via flash silica gel chromatography (5:1 Hex/EtOAc) to give 929 mg of pure product (64% yield). Characterization of the compound agreed with literature data.[4]



(S5) 7-(diethylamino)coumarin-3-carbohydrazide To a round bottom charged with a stir bar and ethyl 7-(diethylamino)coumarin-3-carboxylate (929 mg, 3.21 mmol, 1 eq.), was added anhydrous EtOH (9 mL). To this stirring solution, was added hydrazine monohydrate (0.622 mL, 12.84 mmol, 4 eq.). The reaction mixture was stirred at room temperature until no further precipitate was observed, at which point it was cooled to 0 ºC and stirred for an additional 15 minutes. The precipitate was collected via vacuum filtration to yield pure product (85% yield). Characterization of the compound agreed with literature data.[5]



III. Spectroscopic Studies

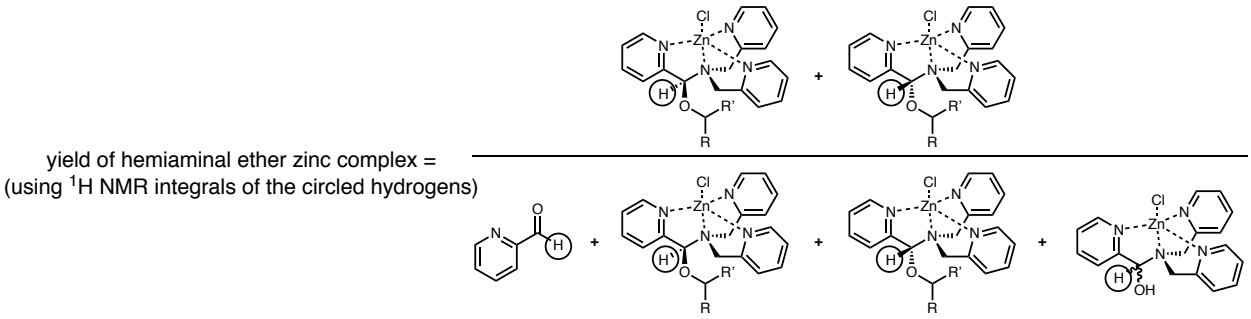
A. Alcohol Assembly (1-position of 2-aminocyclohexanol)

i. Assembly Formation

To a stirred solution of pyridine-2-carboxaldehyde (35 mM, 1 eq.), dipicolylamine (42 mM, 1.2 eq.), 4-(2-chloroethyl)morpholine hydrochloride (35mM, 1 eq.), and zinc (II) triflate (35 mM, 1 eq.), was added *N*-(2-hydroxycyclohexyl)acetamide (175 mM, 5 eq.). Activated 4Å molecular sieves were added and the mixture was stirred overnight at room temperature.

Assembly formation was characterized by ^{1}H-NMR. The percent yield and diastereomeric ratio (*dr*) of the desired hemi-aminal ether product was calculated using integrals from the ^{1}H-NMR. The diastereomeric ratio of the desired hemiaminal ether complex was determined from two peaks in the NMR spectrum in the region of 5.35 – 5.45 ppm for each stereoisomer of *N*-(2-hydroxycyclohexyl)acetamide. The yield of the desired hemiaminal ether complex was determined using the integrals corresponding to the diastereomers of the hemiaminal ether complex, the integral from the hemiaminal complex corresponding to water incorporation, and the integral corresponding to residual pyridine-2-carboxaldehyde (*see below*).

Scheme S1. Yield Determination of Hemiaminal Ether Zinc Complex



(*1R,2R*) dr. = 1.05; yield = 77.7%

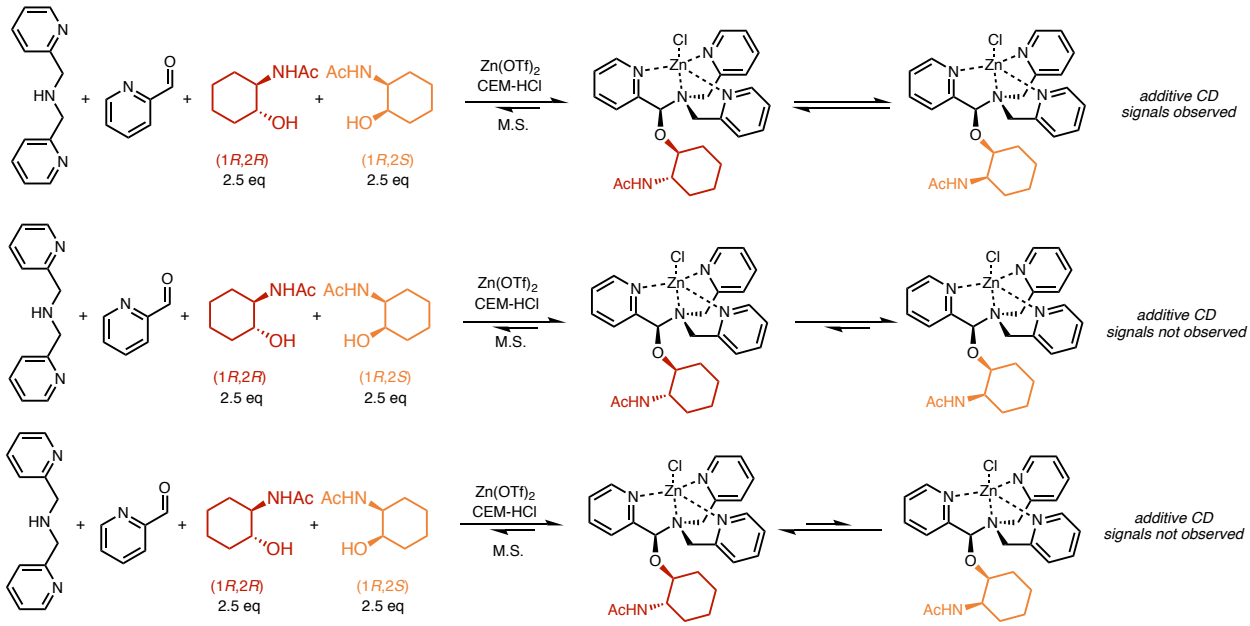
(1R,2S) dr. = 1.32; yield = 74.4%

(1S,2R) dr. = 1.31; yield = 75%

(1S,2S) dr. = 1.11 ; yield = 81.1%

Thus, the average yield of the desired hemiaminal ether zinc complex for the four stereoisomers was determined to be 77.1%. The range of values 74.4 – 81.1% was concluded to be the result of the approximate integrations that were made in the ^{1}H-NMR due to the coincidental overlap between the hemiaminal complex corresponding to water incorporation and one of the diastereomers of the desired hemiaminal ether zinc complex.

Scheme S2. Equilibria expressions showing effects of statistical versus non-statistical alcohol incorporation.



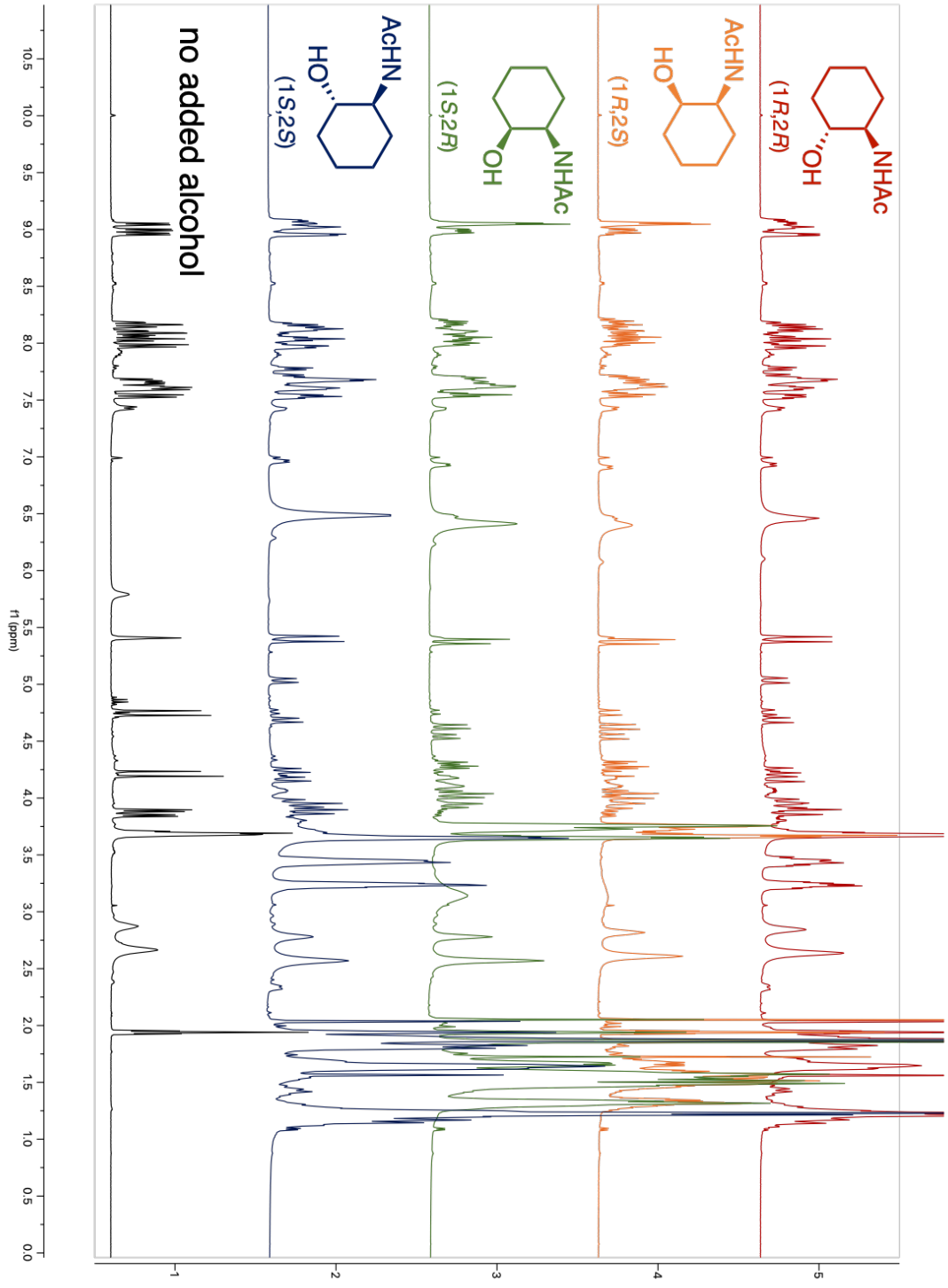


Figure S1. Full NMR spectra for the multicomponent zinc assembly with the four stereoisomers of *N*-(2-hydroxycyclohexyl)acetamide. As required by the principles of stereochemistry, the two sets of enantiomers give identical NMR spectra.

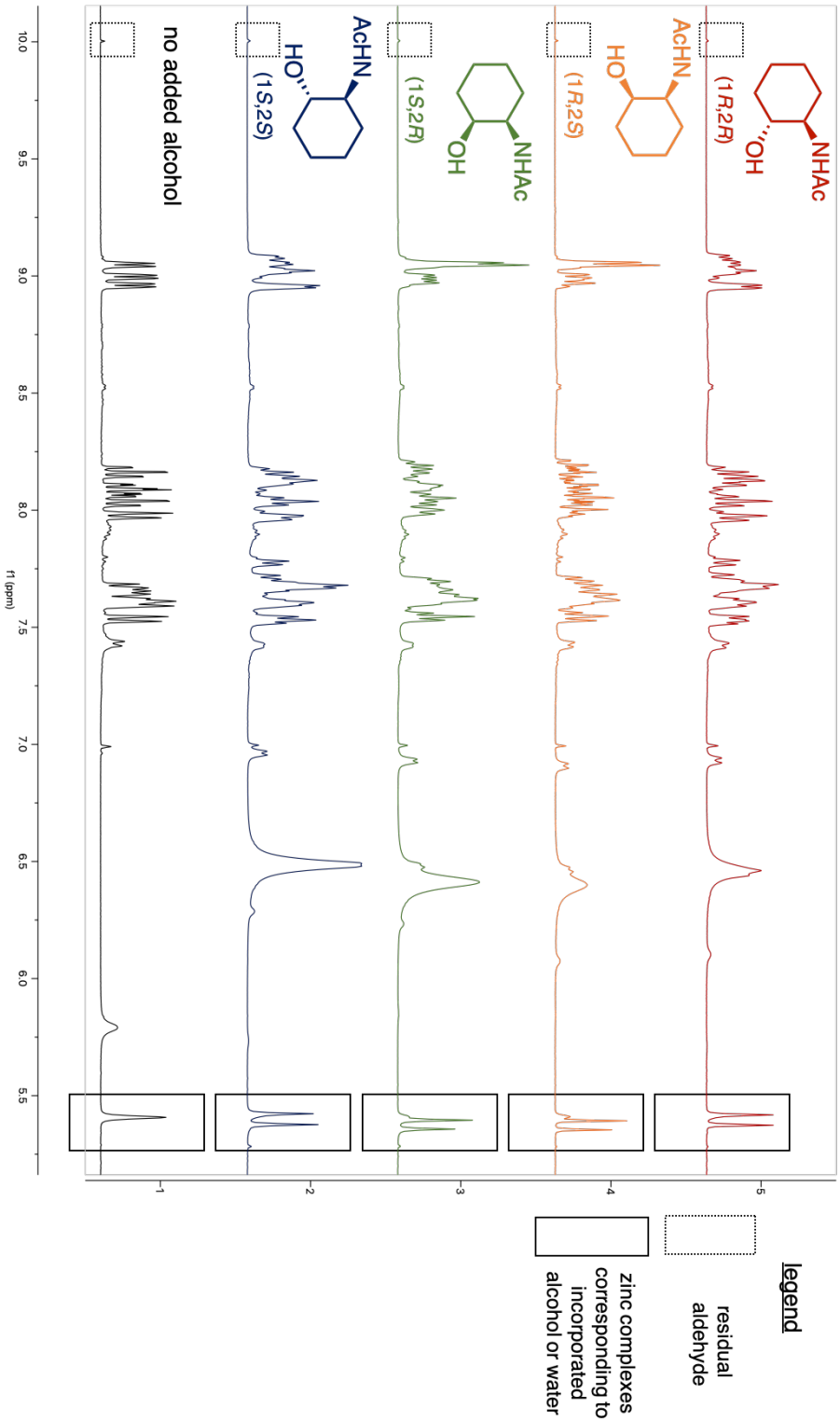


Figure S2. Condensed NMR spectra (5 – 10 ppm) for the multicomponent zinc assembly with the four stereoisomers of *N*-(2-hydroxycyclohexyl)acetamide, with the hydrogen of the residual aldehyde and the hydrogens of the zinc complexes corresponding to incorporated alcohol/water labeled according to the legend.

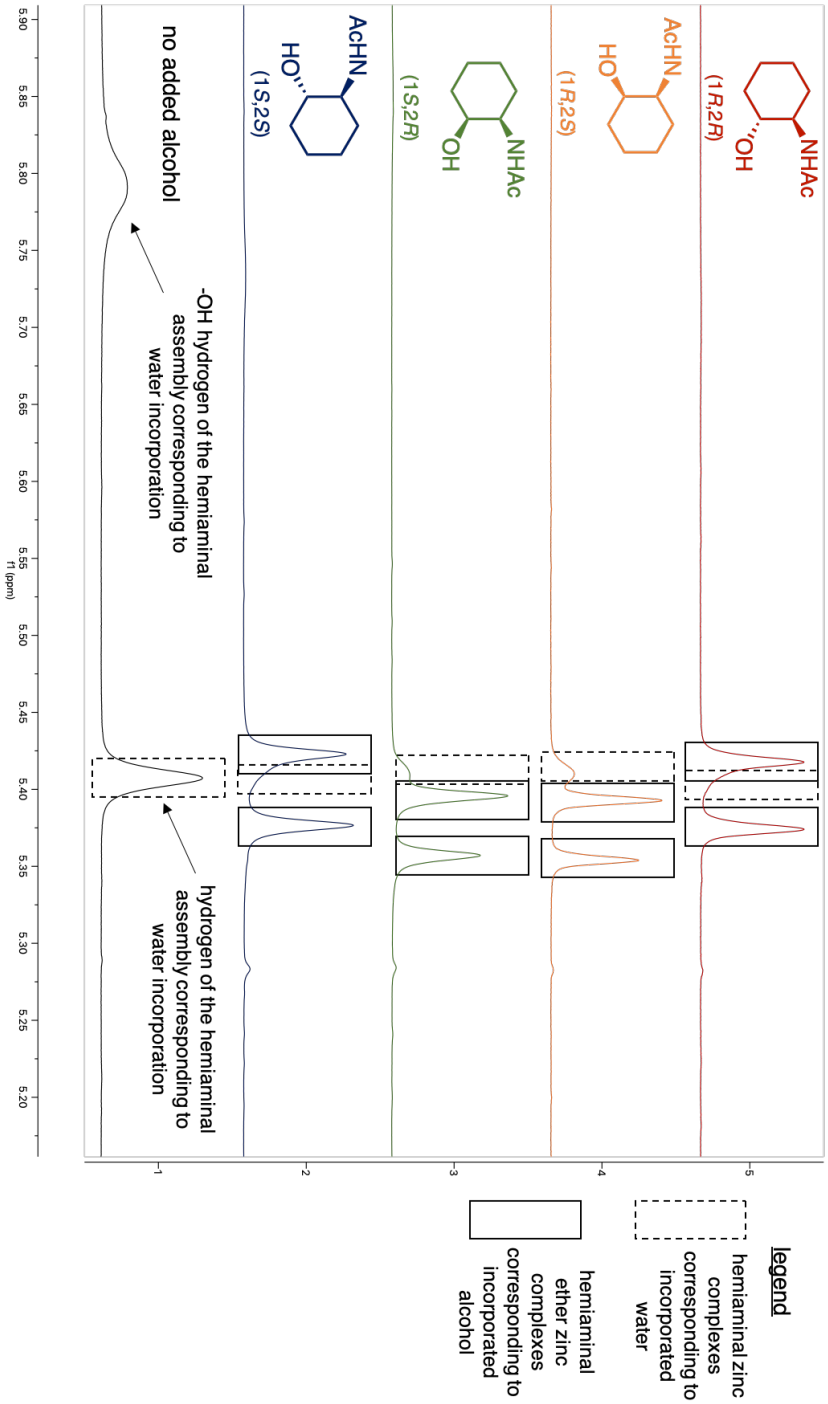


Figure S3. Condensed NMR spectra (5 – 6 ppm) for the multicomponent zinc assembly with the four stereoisomers of *N*-(2-hydroxycyclohexyl)acetamide, where the hydrogens of the zinc complexes corresponding to incorporated alcohol/water are labeled according to the legend. Notice there is overlap between the hemiaminal complex (incorporated water) and the two diastereomers of the hemiaminal ether complex (incorporated alcohol). Approximate integrations were made for the overlapping peaks to calculate diastereomeric ratios and yields for the four stereoisomers of *N*-(2-hydroxycyclohexyl)acetamide.

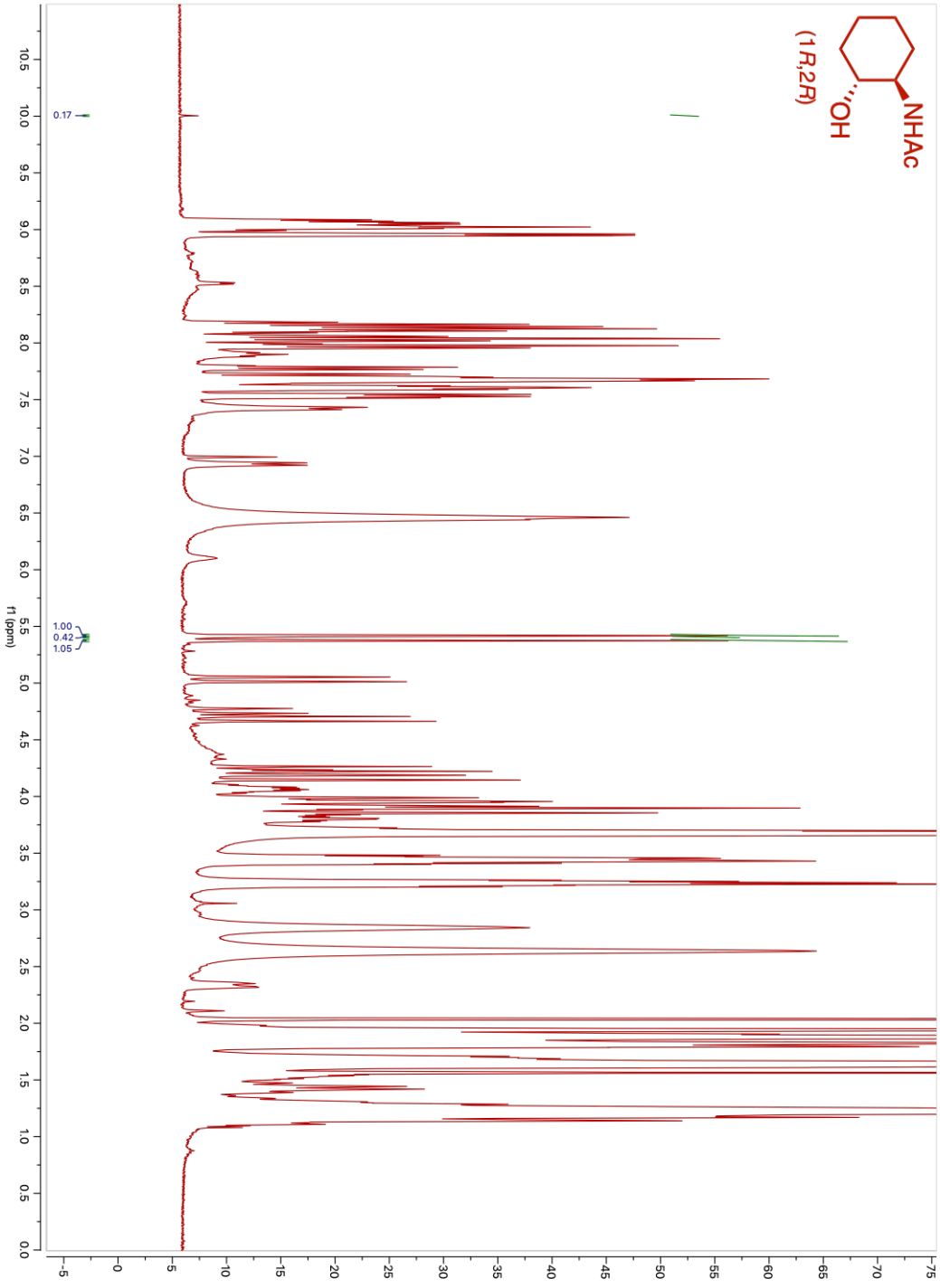


Figure S4. Full NMR spectra for the multicomponent zinc assembly with the (*1R,2R*) isomer of *N*-(2-hydroxycyclohexyl)acetamide, where the integrals relevant for yield determination of the hemiaminal ether zinc complex and diastereomeric ratio for the newly formed stereocenter are shown.

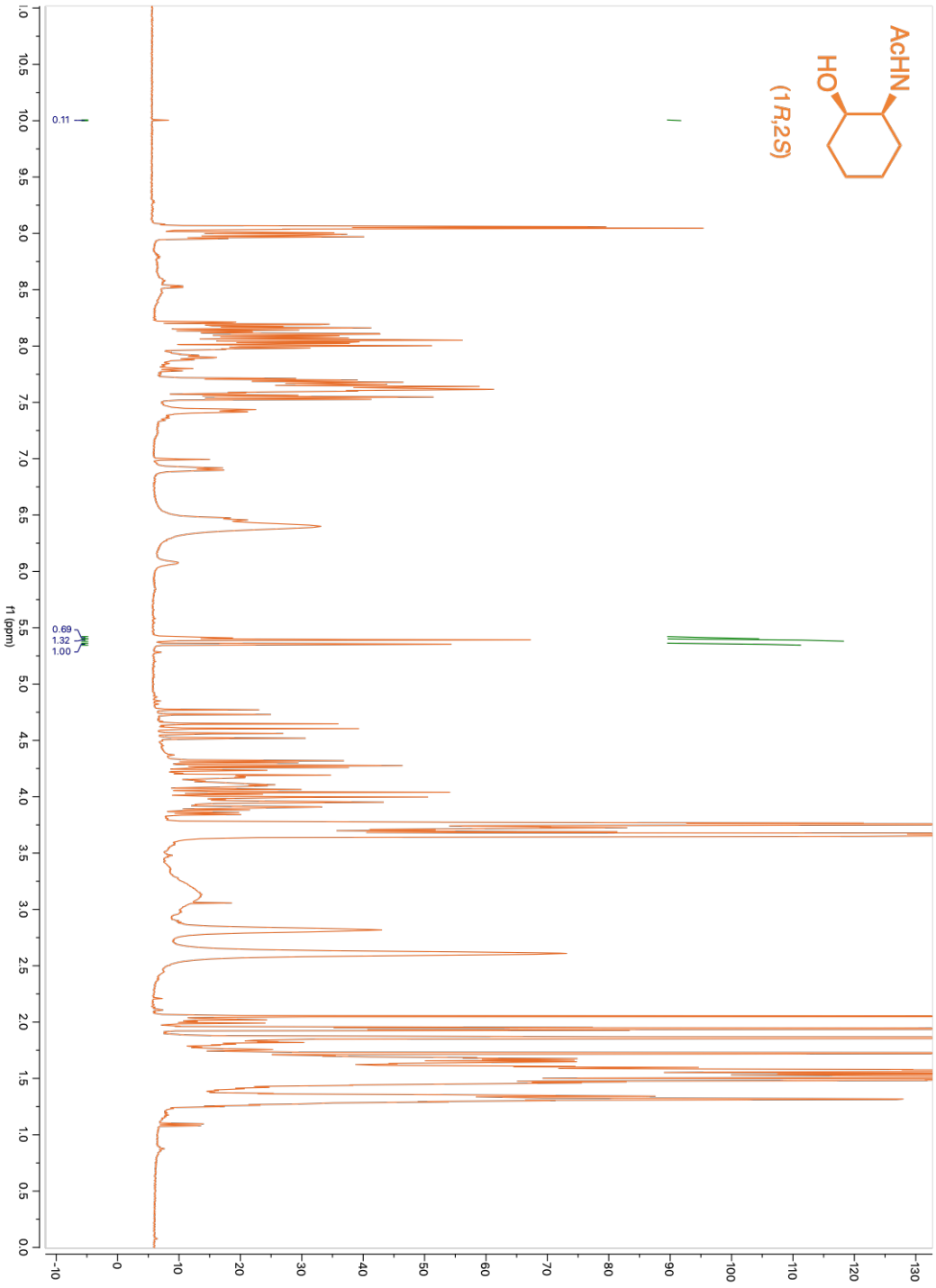


Figure S5. Full NMR spectra for the multicomponent zinc assembly with the (*1R,2S*) isomer of *N*-(2-hydroxycyclohexyl)acetamide, where the integrals relevant for yield determination of the hemiaminal ether zinc complex and diastereomeric ratio for the newly formed stereocenter are shown.

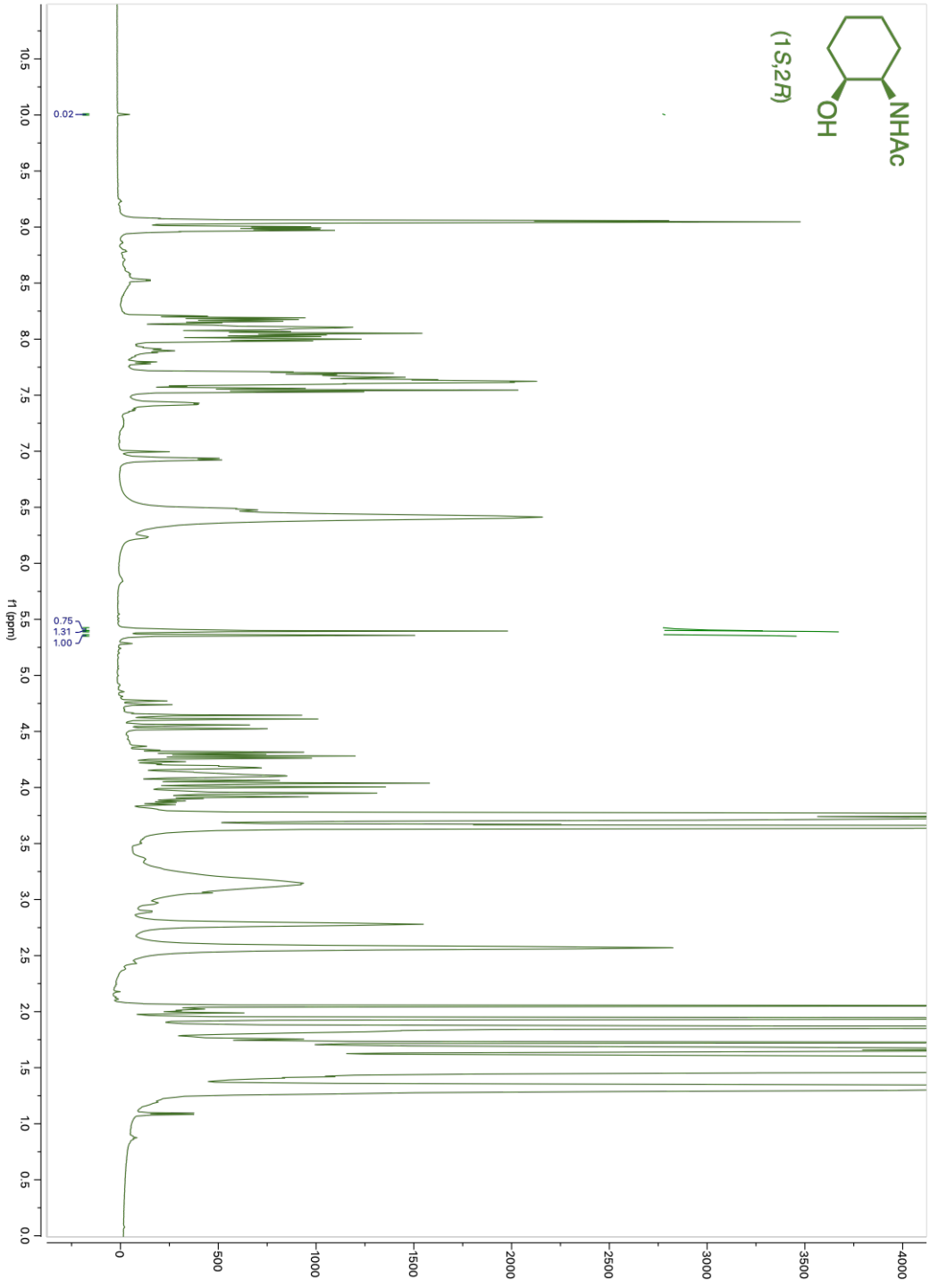


Figure S6. Full NMR spectra for the multicomponent zinc assembly with the (*1S,2R*) isomer of *N*-(2-hydroxycyclohexyl)acetamide, where the integrals relevant for yield determination of the hemiaminal ether zinc complex and diastereomeric ratio for the newly formed stereocenter are shown.

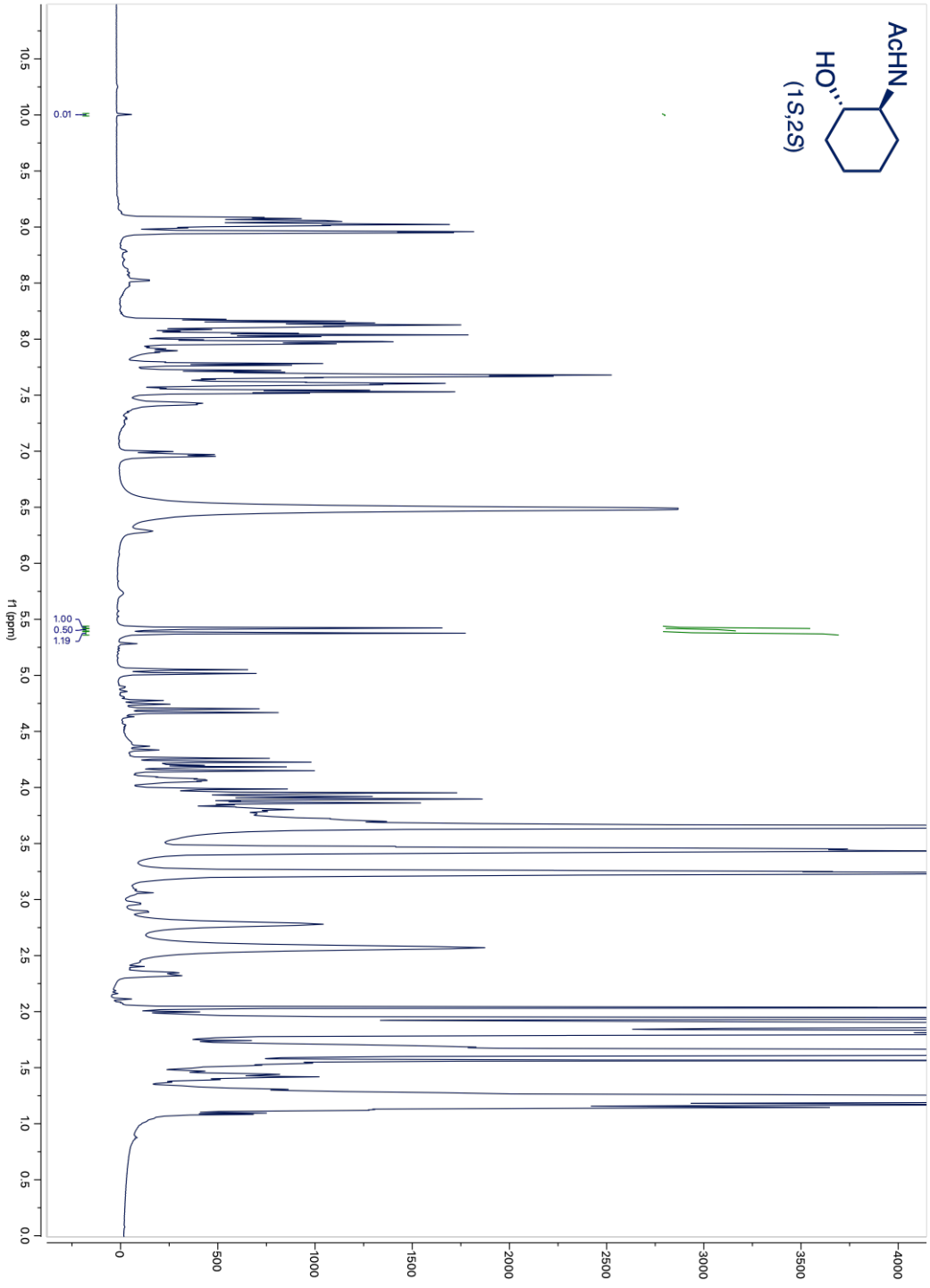


Figure S7. Full NMR spectra for the multicomponent zinc assembly with the (*1S,2S*) isomer of *N*-(2-hydroxycyclohexyl)acetamide, where the integrals relevant for yield determination of the hemiaminal ether zinc complex and diastereomeric ratio for the newly formed stereocenter are shown.

ii. CD Analysis

The CD spectra were gathered using diluted solutions of assembly at 25 ºC (1.75 mM 2-pyridinecarboxaldehyde, 1 mm cell). CD spectra were recorded from 220 – 300 nm with a scan speed of 200 nm/min, and 1 second response time. The CD spectra for each sample were accumulated 3 times.

iii. Enantiomeric Differentiation

For each diastereomeric set of enantiomers, the CD spectra were gathered for solutions of various *ee* (-100, -80, -60, -40, -20, 0, 20, 40, 60, 80, and 100%). The CD intensity at 267 nm was plotted versus *ee* and linear calibration curves were constructed for both diastereomeric sets.

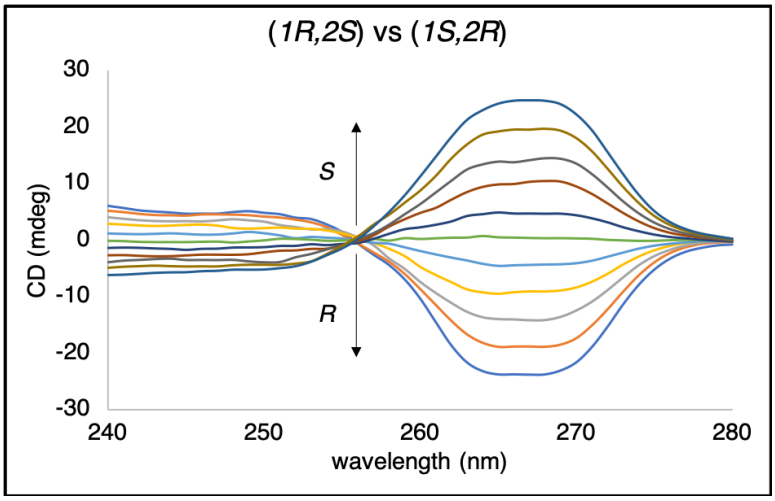


Figure S8. CD spectra of the multicomponent zinc assembly for the *cis*- diastereomeric set of *N*-(2-hydroxycyclohexyl)acetamide with different *ee* values (Top to bottom -100, -80, -60, -40, -20, 0, 20, 40, 60, 80, 100).

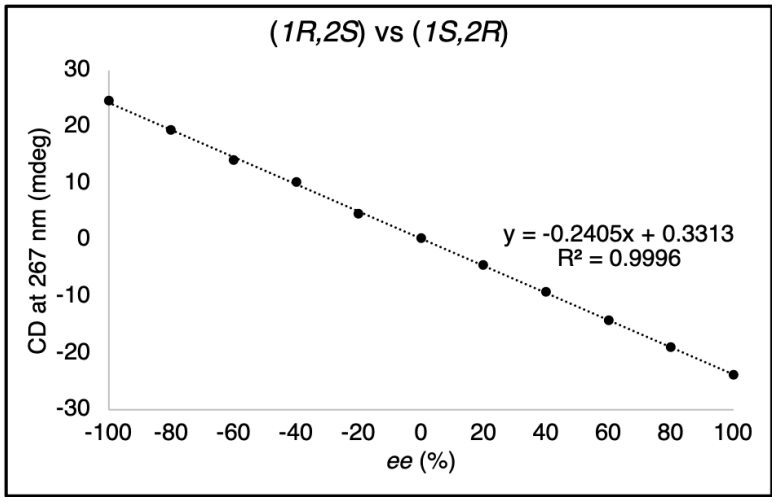


Figure S9. Calibration curve constructed for the *cis*- diastereomeric set of *N*-(2-hydroxycyclohexyl)acetamide that correlates an *ee* value to a CD intensity at 267 nm.

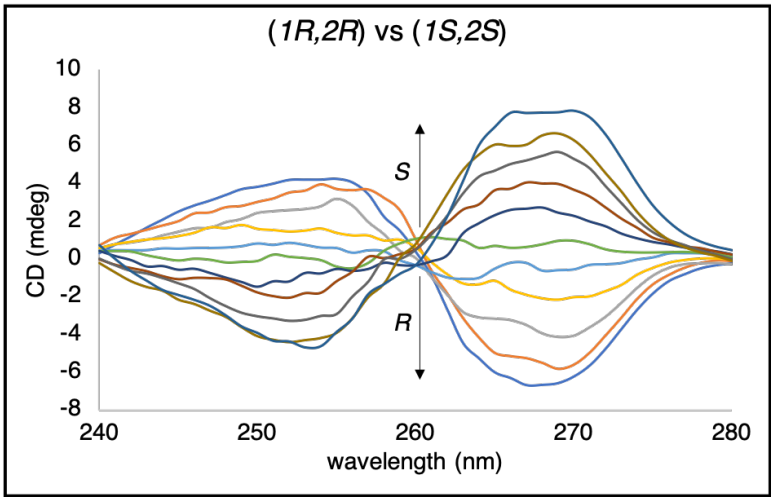


Figure S10. CD spectra of the multicomponent zinc assembly for the *trans*- diastereomeric set of *N*-(2-hydroxycyclohexyl)acetamide with different *ee* values (Top to bottom -100, -80, -60, -40, -20, 0, 20, 40, 60, 80, 100).

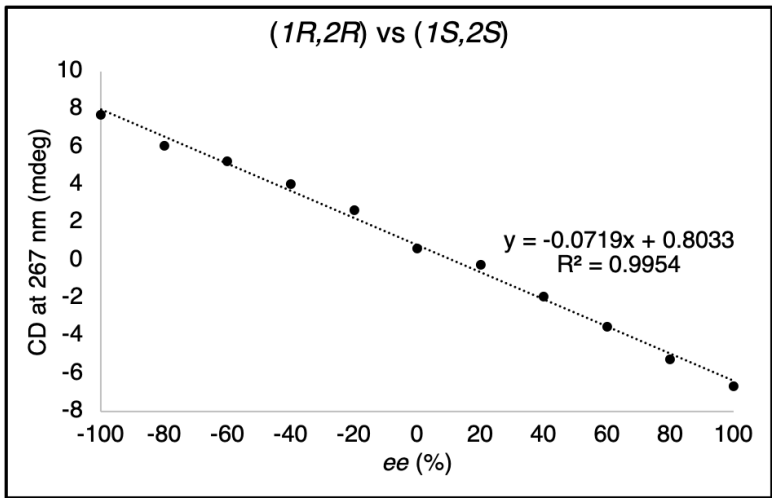


Figure S11. Calibration curve constructed for the *trans*- diastereomeric set of *N*-(2-hydroxycyclohexyl)acetamide that correlates an *ee* value to a CD intensity at 267 nm.

iv. Competition Study Between Epimeric Alcohols

The multicomponent zinc assembly requires five equivalents of the alcohol analyte to ensure efficient incorporation of the alcohol, and give CD spectra that are concentration independent. With the use of five equivalents of alcohol, there is the possibility that one diastereomer of *N*-(2-hydroxycyclohexyl)acetamide may preferentially be incorporated into the zinc assembly. In this case, the observed CD signal would not be a faithful representation of the stereoisomeric mixture because the preferred diastereomer of *N*-(2-hydroxycyclohexyl)acetamide would dominate the signal and any resulting ee_1 determinations would not be correct. For enantiopure substrates, ^{1}H-NMR integrals were used to determine each stereoisomer was incorporated to the same extent,

with an average yield of 77%. If one of the hemiaminal ether zinc complexes derived from either *cis*- or *trans*-*N*-(2-hydroxycyclohexyl)acetamide is significantly more stable than that of its diastereomer, one of the diastereomers will not contribute to the observed CD signal, resulting in incorrect ee_I determinations. If, however, the diastereomers of *N*-(2-hydroxycyclohexyl)acetamide are incorporated on a statistical basis, the observed CD signal will accurately reflect the composition of the stereoisomeric mixture. As CD is a function of absorbance, it can be treated analogously as a UV-Vis spectra, which is the simply sum of the absorbances (or CD) of the individual components. To determine if diastereomers of the analyte were incorporated into the assembly on a statistical basis, 1:1 diastereomeric mixtures were made with 5 equivalents of alcohol e.g. 2.5 eq: 2.5 eq (*1R,2R*):(*1R,2S*) and (*1S,2S*):(*1S,2R*) and subjected to the reaction conditions for alcohol assembly formation. Upon analysis, it was observed that the observed CD signal for the 1:1 diastereomeric mixtures were indeed additive. This was confirmed by analyzing solutions of a 1:1 diastereomeric mixture that were made up from hemiaminal ether zinc complexes derived from enantiopure analytes that gave the same CD signals (*vide infra*). The hemiaminal ether zinc complexes have been reported to exchange incorporate alcohol; thus, to prevent the possible exchange of diastereomers, the mixtures were made at concentrations suitable for chiroptical analysis, rather than the concentrations for assembly formation.

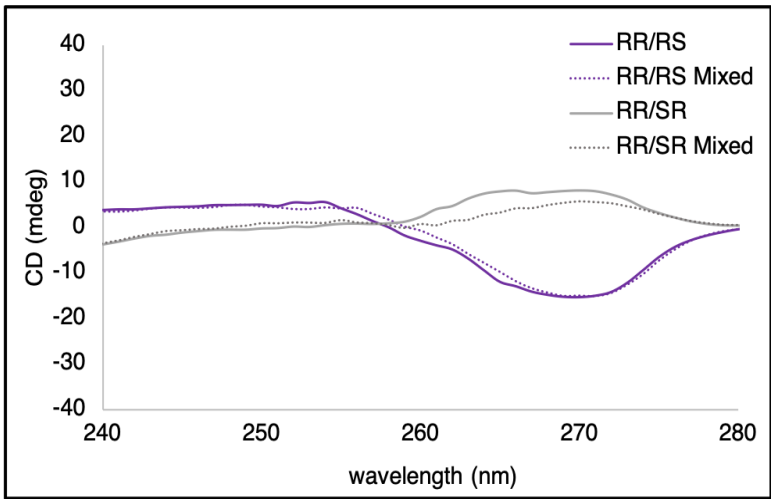


Figure S12. CD spectra for 1:1 competition studies between epimeric alcohols i.e (*1R,2R*):(*1R,2S*) and (*1R,2R*):(*1S,2R*) . The solid lines (*RR/RS* and *RR/SR*) correspond to the 1:1 competition mixtures where there were 2.5 eq of each diastereomer in the reaction mixture, whereas the dotted lines (*RR/RS* Mixed and *RR/SR* Mixed) correspond to the 1:1 mixtures that were made from enantiopure assembly solutions.

B. Amine Assembly (2-position of 2-aminocyclohexanol)

i. Assembly Formation

To a stirred solution of 3-hydroxypyridine-2-carboxaldehyde (15 mM, 3 eq.) and 2-aminocyclohexanol (15 mM, 3 eq.), was added iron (II) triflate (5 mM, 1 eq.) and the solution was incubated for 5 hours.

ii. CD Analysis

The CD spectra were gathered using diluted solutions of assembly at 25 ºC (0.5 mM $Fe(OTf)_2$, 1 cm cell). CD spectra were recorded from 400 – 700 nm with a scan speed of 80 nm/min, and a 1 second response time.

iii. UV-Vis Analysis

The UV-Vis spectra were gathered using diluted solutions of assembly at 25 ºC (0.5 mM $Fe(OTf)_2$, 1 cm cell). UV-Vis spectra were recorded from 400 – 700 nm with a scan speed of 80/nm min, and a 1 second response time.

iv. Enantiomeric Differentiation

For each diastereomeric set of enantiomers, the CD spectra were gathered for solutions of various *ee* (-100, -80, -60, -40, -20, 0, 20, 40, 60, 80, and 100%). The CD intensity at 583 nm was plotted versus *ee* and linear calibration curves were constructed for both diastereomeric sets.

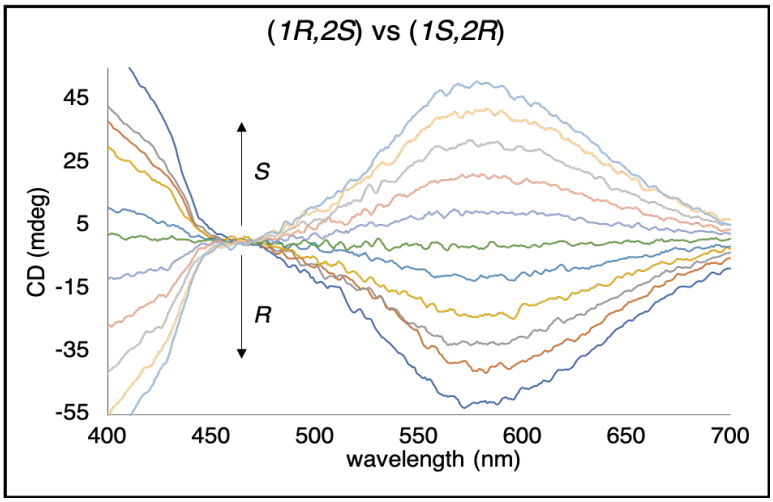


Figure S13. CD spectra of the octahedral iron complexes for the *cis*- diastereomeric set of 2-aminocyclohexanol with different *ee* values (Top to bottom -100, -80, -60, -40, -20, 0, 20, 40, 60, 80, 100).

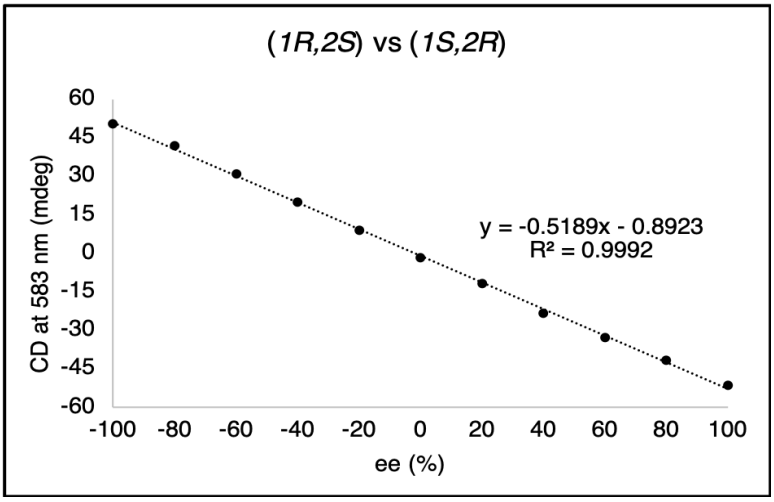


Figure S14. Calibration curve constructed for the *cis*- diastereomeric set of 2-aminocyclohexanol that correlates an *ee* value to a CD intensity at 583 nm.

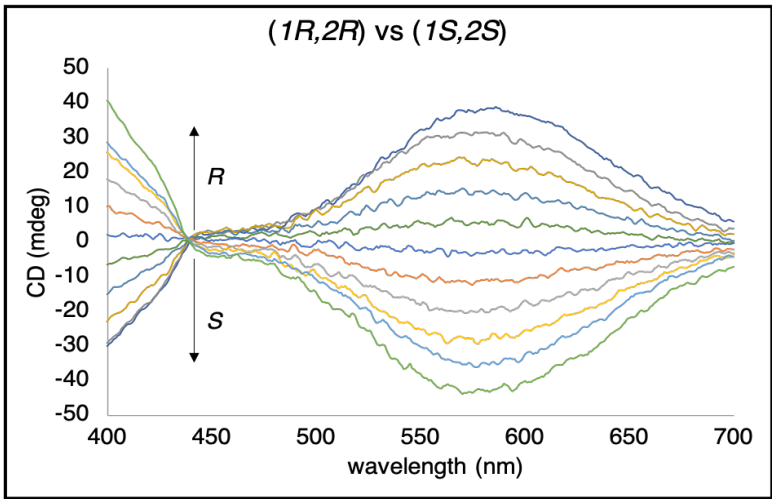


Figure S15. CD spectra of the octahedral iron complexes for the *trans*- diastereomeric set of 2-aminocyclohexanol with different *ee* values (Top to bottom 100, 80, 60, 40, 20, 0, -20, -40, -60, -80, -100).

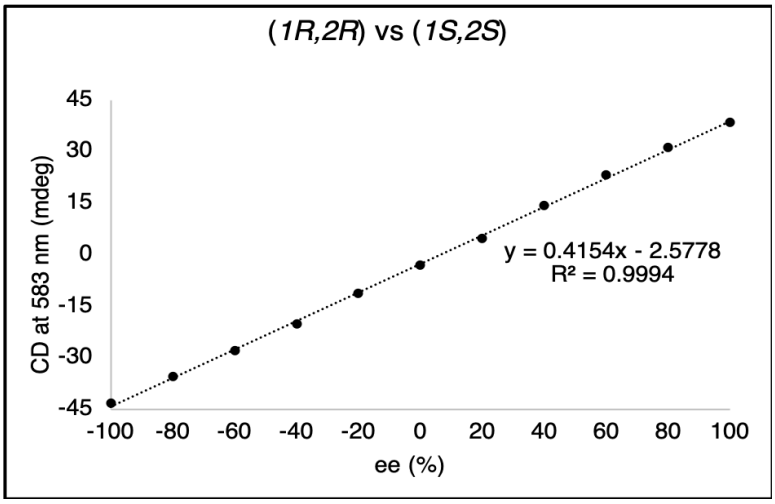


Figure S16. Calibration curve constructed for the *trans*- diastereomeric set of 2-aminocyclohexanol that correlates an *ee* value to a CD intensity at 583 nm.

v. Diastereomeric Differentiation

For each diastereomeric combination of 2-aminocyclohexanol i.e. (*1R,2R*) vs (*1R,2S*), (*1R,2R*) vs (*1S,2R*), (*1S,2S*) vs (*1R,2S*) and (*1S,2S*) vs (*1S,2R*)of 2-aminocyclohexanol, the absorbance spectra were gathered for solutions of various *de* (-100, -80, -60, -40, -20, 0, 20, 40, 60, 80, and 100%). The intensity at 490 nm was plotted versus *de* and linear calibration curves were constructed for both diastereomeric sets. A titration was also performed for each pair of enantiomers of 2-aminocyclohexanol i.e. (*1R,2R*) vs (*1S,2S*), and (1R,2S) vs (1S,2R). The absorbance spectra were gathered for solutions of various *ee* (-100, -80, -60, -40, -20, 0, 20, 40, 60, 80, and 100%). The intensity at 490 nm was plotted versus *ee* and there was no appreciable difference in absorbance intensity at 490 nm of the solutions with varied *ee*.

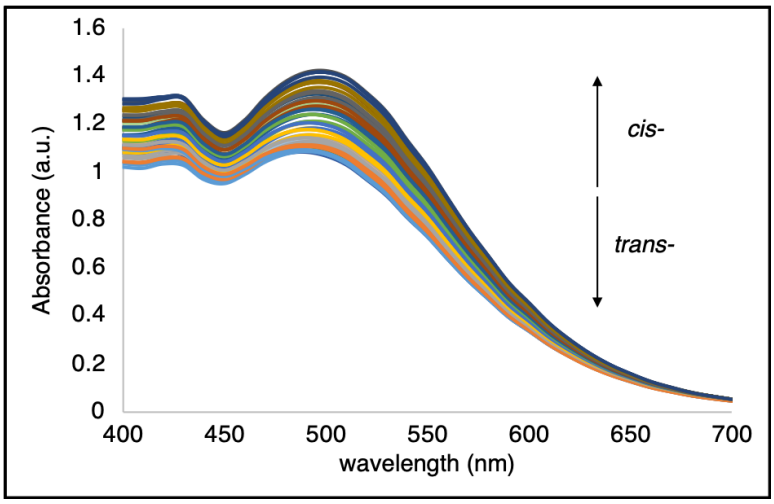


Figure S17. Overlaid UV-Vis spectra of the octahedral iron complexes with different *de* values (Top to bottom 100, 80, 60, 40, 20, 0, -20, -40, -60, -80, -100) for every diastereomeric combination of 2-aminocyclohexanol.

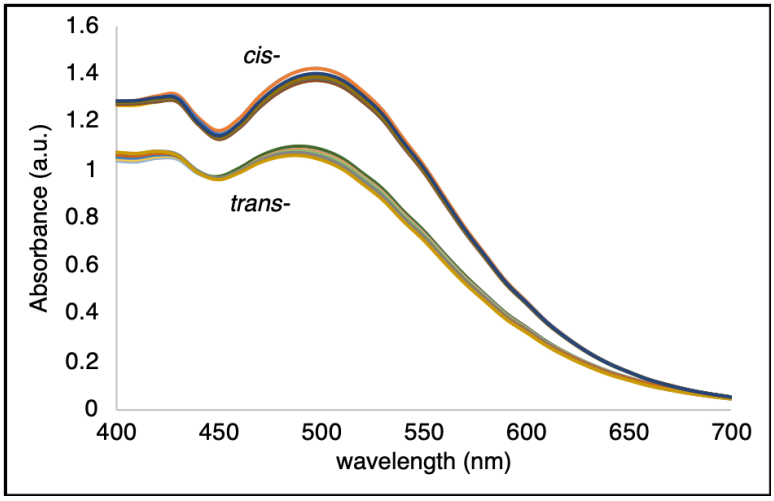


Figure S18. UV-Vis spectra of the octahedral iron complexes for the *cis*- ((*1R,2S*) vs (*1S,2R*)) and *trans*- ((*1R,2R*) vs (*1S,2S*)) enantiomeric pairs with different *ee* values (100, 80, 60, 40, 20, 0, -20, -40, -60, -80, -100).

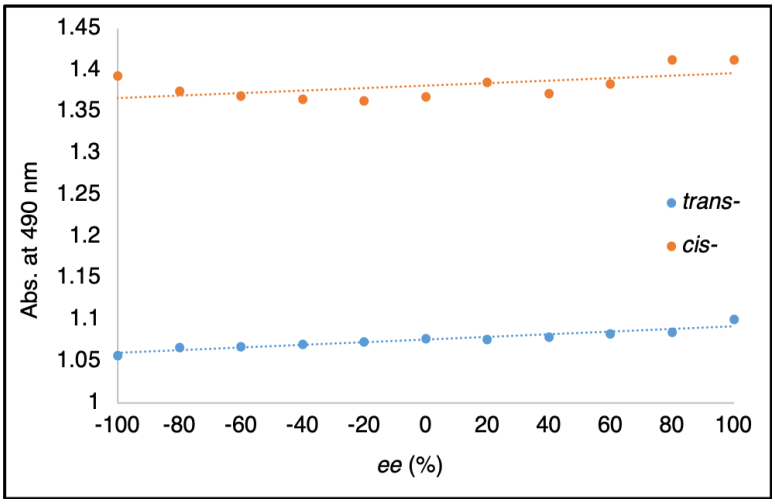


Figure S19. Absorbance intensity of the octahedral iron complexes at 490 nm plotted versus *ee* for the *cis*- ((*1R,2S*) vs (*1S,2R*)) and *trans*- ((*1R,2R*) vs (*1S,2S*)) enantiomeric pairs.

vi. Competition Studies Between Epimeric Amines

The octahedral iron complexes are formed at a ratio of 3:3:1 2-aminocyclohexanol, 3-hydroxypyridine-2-carboxaldehyde and iron (II) triflate. For an enantiopure amine, there are 4 stereoisomers of the octahedral iron complex as a result of the handedness of the amine (*R* and *S*), the helicity of the octahedral iron complex (Λ- and Δ-), and the configurational isomerism of the octahedral iron complex (*fac*- and *mer*-). For a scalemic mixture, there are 28 stereoisomers of the octahedral iron complex, again due to the three types of isomerism associated with the octahedral iron complex. This stereochemical complexity does not affect chiroptical analysis because the 28 stereoisomers rapidly interconvert in equilibria. In the context of a chiral compound with four stereoisomers, the system becomes even more stereochemically complex. In order for the chiroptical analyses to be unaffected by this stereochemical complexity, the chiral imine ligands should dock at the iron (II) center on a purely statistical basis; any cooperativity among ligands would affect the observed CD signal and the resulting ee_2 determination would be incorrect. To determine if diastereomers of the analyte were incorporated into the assembly on a statistical basis, 1:1 diastereomeric mixtures were made with 3 equivalents of amine e.g. 1.5 eq: 1.5 eq (*1R,2R*):(*1R,2S*) and (*1S,2S*):(*1S,2R*) and subjected to the reaction conditions for amine assembly formation. As CD is a function of absorbance, it can be treated analogously as a UV/Vis spectra, which is simply the sum of the absorbances (or CD) of the individual components. In our case, it was observed that the observed CD signal for the 1:1 diastereomeric mixtures were indeed additive. This was further confirmed by analyzing a 1:1 diastereomeric mixtures that were made up from two enantiopure solutions that gave the CD signals with no significant differences (*vide infra*).

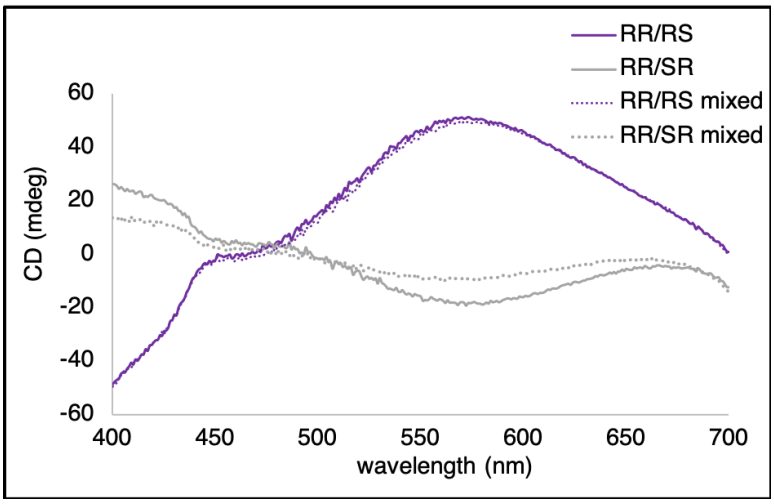


Figure S20. CD spectra of the octahedral iron complexes for 1:1 competition studies between epimeric amines i.e (*1R,2R*):(*1R,2S*) and (*1R,2R*):(*1S,2R*) . The solid lines (*RR/RS* and *RR/SR*) correspond to the 1:1 competition mixtures with 1.5 eq of each diastereomer, whereas the dotted lines (*RR/RS* mixed and *RR/SR* mixed) correspond to the 1:1 mixtures that were made from enantiopure assembly solutions.

C. Concentration Determination

A fluorescence-based concentration assay was designed using a fluorescent coumarin acyl hydrazide (**S5).** Condensation of the coumarin fluorophore on *N*-(2-oxocyclohexyl)acetamide gave a turn on in fluorescence that could be correlated to the concentration of *N*-(2-oxocyclohexyl)acetamide and thus, percent yield.

To a 10 mM solution of **S5** in acetonitrile was added varying amounts of *N*-(2-oxocyclohexyl)acetamide, and a small scoop of $MgSO_4$. The reactions were incubated at 40 ºC for 3 hours, diluted to 6 μM in acetonitrile and fluorescence spectra were gathered (λ_{exc} = 415 nm; $\lambda_{emission}$ = 470 nm). The emission at 470 nm was plotted against percent yield and a calibration curve was constructed that correlated a fluorescence intensity to a percent yield.

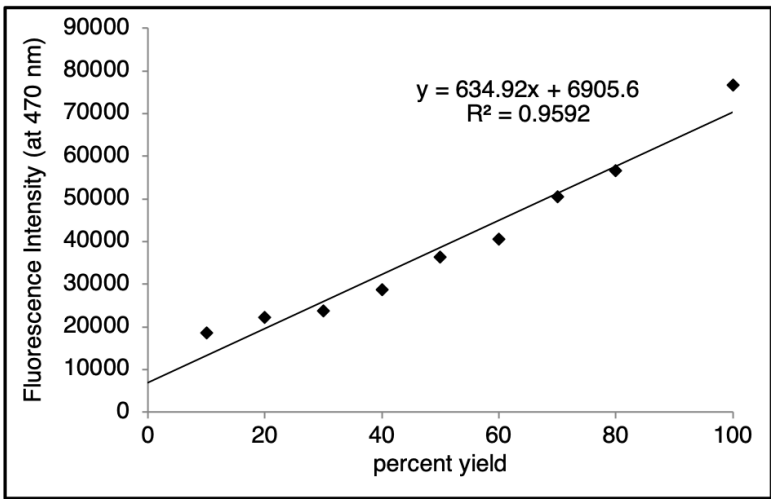


Figure S21. Calibration curve constructed for the concentration determination of *N*-(2-oxocyclohexyl)acetamide

IV. Chemometric Analysis

A. Training Set

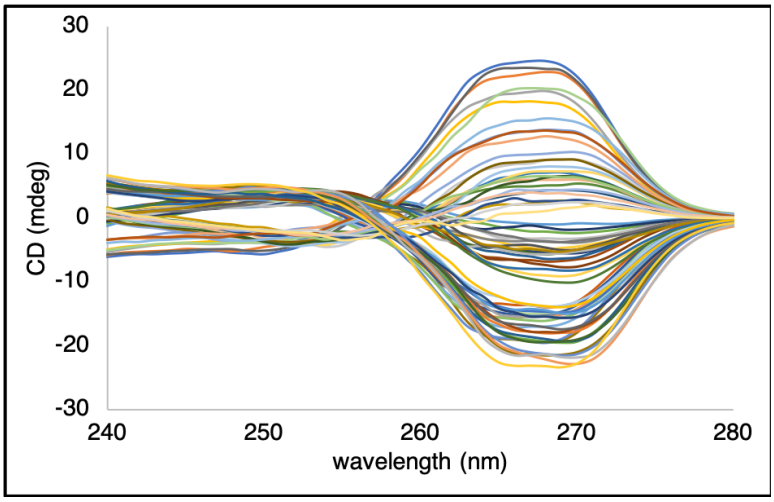


Figure S22. Training set CD spectra using a multicomponent zinc assembly for chiroptical analysis of the 1-position (ee_1) for stereoisomeric mixtures of *N*-(2-hydroxycyclohexyl)acetamide.

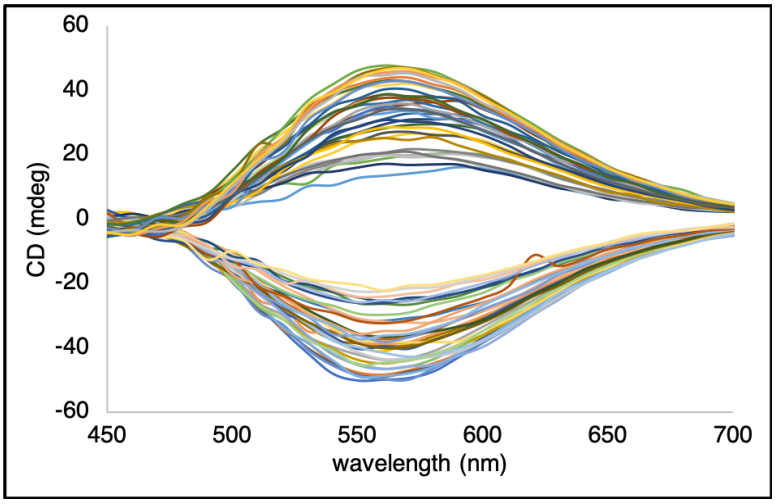


Figure S23. Training set CD spectra using octahedral iron (II) complexes for chiroptical analysis of the 2-position (ee_2) for stereoisomeric mixtures of 2-aminocyclohexanol.

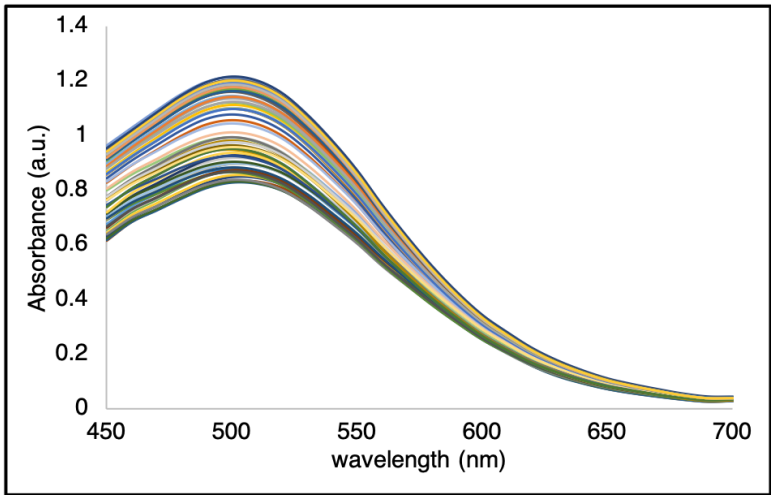


Figure S24. Training set UV-Vis spectra using octahedral iron (II) complexes for analysis of the relative stereochemistry (*de*) for stereoisomeric mixtures of 2-aminocyclohexanol.

B. Test Set Data

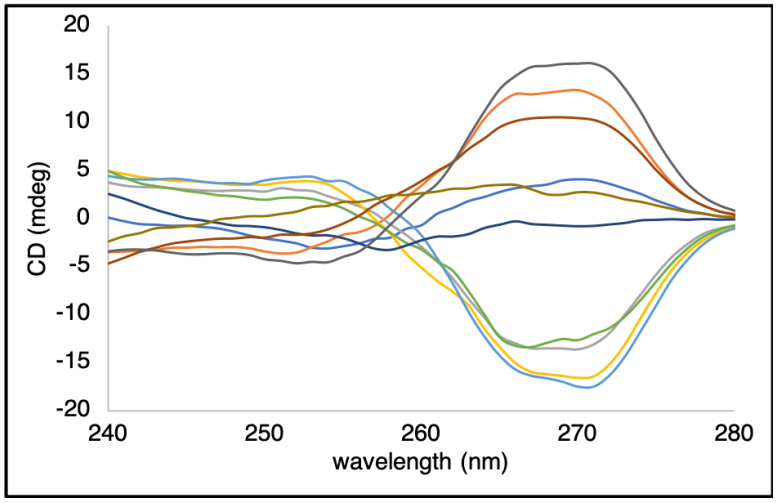


Figure S25. Test set CD spectra using a multicomponent zinc assembly for chiroptical analysis of the 1-position (ee_1) for stereoisomeric mixtures of *N*-(2-hydroxycyclohexyl)acetamide.

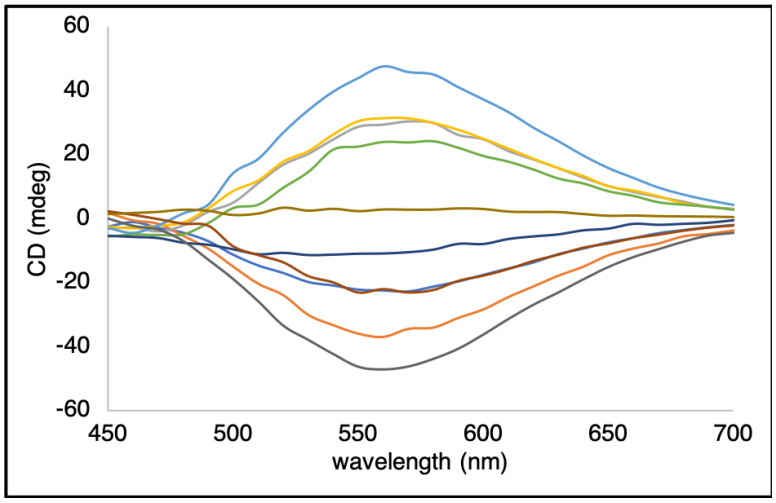


Figure S26. Test set CD spectra using octahedral iron (II) complexes for chiroptical analysis of the 2-position (ee_2) for stereoisomeric mixtures of 2-aminocyclohexanol.

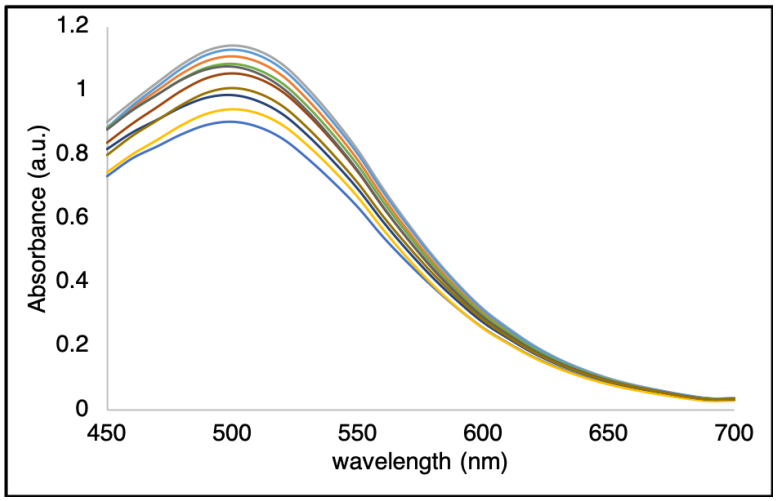


Figure S27. Test set UV-Vis spectra using octahedral iron (II) complexes for analysis of the relative stereochemistry (*de*) for stereoisomeric mixtures of 2-aminocyclohexanol.

C. Determination of ee_1, ee_2 and *de*

Table S1. Determination of ee_1, ee_2 using SO-CovSel and *de* using the absorbance intensity at 490 nm.

ee_1		ee_2		*de*	
calculated	***actual***	***calculated***	***actual***	***calculated***	***actual***
86.38	100	-106.33	-100	-95.64	-100
-96.82	-100	-96.28	-100	81.97	100
-86.21	-90	98.49	100	-79.65	-90
87.56	90	95.21	90	123.93	100
-43.39	-45	-51.17	-45	75.36	75
59.06	60	-50.81	-50	-62.05	-50
31.23	30	101.32	90	16.01	20
-91.34	-100	14.23	20	-27.31	-20
-57.13	-60	-55.88	-60	-23.76	20
67.30	60	-68.10	-70	-35.66	-30

V. Detailed Description of Sequential and Orthogonalized Covariance Selection (SO-CovSel)

When a single set of measurements (collected in the data matrix **X**) are recorded on a system with the aim of predicting one or more responses (**Y**), their mathematical relation is described by a (usually linear) regression model relating **X** to **Y** through:

$$\boldsymbol{Y = XB}$$

B being a set of coefficients fully determining the relation and whose optimal values are estimated during the model building stage (e.g., by a least-squares approach).

On the other hand, when different sets of measurements are collected on the same samples, such as in the present study, where
the CD intensities for the alcohol ($\mathbf{X}_{CD,\ alcohol}$) and for the amine ($\mathbf{X}_{CD,\ amine}$) moieties, and the UV-Vis spectra of the amine ($\mathbf{X}_{UV\text{-}Vis,\ amine}$) were recorded, instead of collecting all the variables in a single matrix

$$\boldsymbol{X} = [\boldsymbol{X}_{CD,alcohol} \quad \boldsymbol{X}_{CD,amine} \quad \boldsymbol{X}_{UV-Vis,amine}]$$

it is advisable to use chemometric methods specifically designed to deal with such data configuration, which are called multi-block approaches. In the calibration framework, a multi-block regression model can be expressed by the general formula:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{B}_1 + \boldsymbol{X}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{X}_k\boldsymbol{B}_k$$

k being the number of blocks (three, in the present study), and each of the matrices $\boldsymbol{B}_i$ ($\boldsymbol{B}_1$, $\boldsymbol{B}_2$, …, $\boldsymbol{B}_k$) collecting the coefficients relating the corresponding block to the responses. Among the possible multi-block approaches available in the literature, in the present study it was chosen to use a recently proposed method called Sequential and Orthogonalized Covariance Selection (SO-CovSel), due to its characteristics which are particularly suited for designing sensors. Indeed, as the name suggests, SO-CovSel adopts a sequential modeling scheme in which the blocks are modeled one after another and orthogonalization allows that only new information is modeled at each successive step. Moreover, by coupling this multi-block regression scheme with variable selection according to the Covariance Selection method[6] only a small number of variables for each block are used to build the calibration model, which is then more easily interpretable, more robust and "parsimonious".
In detail, considering a general case involving, as in the present study, three blocks of predictors ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$) to predict the responses $\boldsymbol{Y}$, the SO-CovSel algorithm is based on the following steps, alternating variable selection and regression:

1. The blocks $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are mean centered and $\boldsymbol{Y}$ is auto-scaled; the number of variables to be selected in each block L_1, L_2 and L_3 is set.
2. The Covariance Selection procedure is used to select the L_1 variables from the first block $\boldsymbol{X}_1$ to obtain $\boldsymbol{X}_{1,sel}$:
    a. The variable $\boldsymbol{x}_{1,sel}$ with the highest covariance with the responses (i.e., the column $\boldsymbol{x}_{1,i}$ of the matrix $\boldsymbol{X}_1$ satisfying $\arg\max_i(\boldsymbol{x}_i^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{x}_i)$) is selected.
    b. $\boldsymbol{X}_1$ is orthogonalized with respect to $\boldsymbol{x}_{1,sel}$ to obtain $\boldsymbol{X}_1^o$: $\boldsymbol{X}_1^o = \boldsymbol{X}_1 - \boldsymbol{x}_{1,sel}\left(\boldsymbol{x}_{1,sel}^T\boldsymbol{x}_{1,sel}\right)^{-1}\boldsymbol{x}_{1,sel}^T\boldsymbol{X}_1$
    c. $\boldsymbol{Y}$ is orthogonalized with respect to $\boldsymbol{x}_{1,sel}$ to obtain $\boldsymbol{Y}^o$: $\boldsymbol{Y}^o = \boldsymbol{Y} - \boldsymbol{x}_{1,sel}\left(\boldsymbol{x}_{1,sel}^T\boldsymbol{x}_{1,sel}\right)^{-1}\boldsymbol{x}_{1,sel}^T\boldsymbol{Y}$
    d. Steps *a* to *c* are repeated substituting $\boldsymbol{X}_1$ with $\boldsymbol{X}_1^o$ and $\boldsymbol{Y}$ with $\boldsymbol{Y}^o$ until L_1 variables are selected. The selected variables are gathered in the matrix $\boldsymbol{X}_{1,sel}$.
3. A regression model is built between $\boldsymbol{Y}$ and $\boldsymbol{X}_{1,sel}$ by ordinary least squares regression: $\boldsymbol{Y} = \boldsymbol{X}_{1,sel}\boldsymbol{B}_1 + \boldsymbol{E}_1$, $\boldsymbol{E}_1$ being the residuals, i.e., the differences between the measured values of the responses and the model predictions based on the first block ($\boldsymbol{E}_1 = \boldsymbol{Y} - \boldsymbol{X}_{1,sel}\boldsymbol{B}_1$).
4. The second block is orthogonalized with respect to the selected variables of the first block: $\boldsymbol{X}_2^{orth} = \boldsymbol{X}_2 - \boldsymbol{X}_{1,sel}\left(\boldsymbol{X}_{1,sel}^T\boldsymbol{X}_{1,sel}\right)^{-1}\boldsymbol{X}_{1,sel}^T\boldsymbol{X}_2$.

5. The CovSel procedure is used to select variables on the orthogonalized second block analogously to what described in step 2, but using $\boldsymbol{E}_1$ as the response matrix. The L_2 variables extracted from the orthogonalized second block are collected in the matrix $\boldsymbol{X}_{2,sel}^{orth}$.
6. A regression model is built between $\boldsymbol{E}_1$ and $\boldsymbol{X}_{2,sel}^{orth}$ by ordinary least squares regression: $\boldsymbol{E}_1 = \boldsymbol{X}_{2,sel}^{orth}\boldsymbol{B}_{2,orth} + \boldsymbol{E}_2$, $\boldsymbol{E}_2$ being the residuals of this second regression step ($\boldsymbol{E}_2 = \boldsymbol{E}_1 - \boldsymbol{X}_{2,sel}^{orth}\boldsymbol{B}_{2,orth}$).
7. The third block is orthogonalized with respect to the selected variables of the first and the second blocks: $\boldsymbol{X}_3^{orth} = \boldsymbol{X}_3 - \boldsymbol{X}_{12,sel}\left(\boldsymbol{X}_{12,sel}^{\boldsymbol{T}}\boldsymbol{X}_{12,sel}\right)^{-1}\boldsymbol{X}_{12,sel}^{\boldsymbol{T}}\boldsymbol{X}_3$, where $\boldsymbol{X}_{12,sel} = \left[\boldsymbol{X}_{1,sel}\boldsymbol{X}_{2,sel}\right]$.
8. The CovSel procedure is used to select variables on the orthogonalized third block analogously to what described in step 2, but using $\boldsymbol{E}_2$ as the response matrix. The L_3 variables extracted from the orthogonalized second block are collected in the matrix $\boldsymbol{X}_{3,sel}^{orth}$.
9. A regression model is built between $\boldsymbol{E}_2$ and $\boldsymbol{X}_{3,sel}^{orth}$ by ordinary least squares regression: $\boldsymbol{E}_2 = \boldsymbol{X}_{3,sel}^{orth}\boldsymbol{B}_{3,orth} + \boldsymbol{E}_3$, $\boldsymbol{E}_3$ being the residuals of this third regression step ($\boldsymbol{E}_3 = \boldsymbol{E}_2 - \boldsymbol{X}_{3,sel}^{orth}\boldsymbol{B}_{3,orth}$).
10. The overall predictive model results from the sum of the three individual regression models in steps 3, 6 and 9 and can be expressed as: $\boldsymbol{Y} = \boldsymbol{X}_{1,sel}\boldsymbol{B}_1 + \boldsymbol{X}_{2,sel}^{orth}\boldsymbol{B}_{2,orth} + \boldsymbol{X}_{3,sel}^{orth}\boldsymbol{B}_{3,orth} + \boldsymbol{E}_3$, where it is evident that the residuals of the last regression step coincide with the residuals of the overall regression model.

Here it is worth to stress that, even if in step 10 the model is expressed in terms of the orthogonalized blocks $\boldsymbol{X}_{2,sel}^{orth}$ and $\boldsymbol{X}_{3,sel}^{orth}$, by doing a bit of algebra it is also possible to directly express it in terms of the original variables.
As it is also evident from the procedure, the final model depends on the number of variables to be selected in each block (L_1, L_2 and L_3): the optimal value of these meta-parameters has to be optimized during the model building stage, usually by means of a cross-validation procedure.

References

1. Mueller, C.E.; Zell, D.; Hrdina, R.; Wende, R. C.; Wanka, L; Schuler, Soeren M.M.; Schreiner, P.R. *JOC*. **2013**, *78*, 17, 8465 – 8484.
2. Schaus, S.; Larrow, J.; Jacobsen, E. *JOC*. **1997**, *62*, 12, 4197-4199.
3. Kim, Y.; Pak, H.K.; *Chem. Commun.*, **2016**, *52*, 6549-6552.
4. Ma Y.; Luo W.; Quinn P.J.; Liu Z.; Hider R.C. *J Med Chem*. **2004**, *47*, 6349-6362.
5. Takechi,H.; Oda Y.; Nishizono N.; Oda K, Machida M. *Chem Pharm Bull.* **2000**, *48*, 1702-1710.
6. Roger, J.M.; Palagos, B.; Bertrand, D.; Fernandez-Ahumada, E., CovSel: variable selection for highly multivariate and multi-response calibration application to IR spectroscopy. *Chemometr. Intell. Lab. Syst.* **2011**, *106(2)*, 216-223.